DYNAMIC RESOURCE ALLOCATION IN MANUFACTURING AND SERVICE INDUSTRIES

A Thesis
Presented to
The Academic Faculty

by

Tuba Yılmaz

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2013

DYNAMIC RESOURCE ALLOCATION IN MANUFACTURING AND SERVICE INDUSTRIES

Approved by:

Professor Pınar Keskinocak, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Esma Gel, Co-advisor
School of Computing, Informatics & Decision Systems Engineering
Arizona State University

Professor David M. Goldsman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Julie Swann
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Alan Erera
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 11 January 2013

To my beloved parents,

Hatice-Muzaffer Yılmaz

and

my brother

Alper Yılmaz

for their unconditional love and support...

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Pinar Keskinocak, for providing me the research opportunity, guidance and support during this study. I would like to express special thanks to my co-advisor, Dr. Esma Gel, for her guidance and valuable insights in my studies. My gratitude also goes to Dr. David Goldsman who encouraged and supported me from the beginning of my graduate studies and introduced me to the ISyE program. I thank Dr. Julie L. Swann and Dr. Alan Erera for their time and effort in serving on my Ph.D. committee.

I would like to thank Dr. Ali Ekici, Dr. Baykal Hafizoglu, Dr. Secil Savasaneril for sharing their experience and giving me guidance. I also would like to take this chance to thank Dr. Ilhan Or and Dr. Birnur Ozbas from Bogazici University with whom I started my first research project and decided to pursue graduate studies with their encouragement and support.

I feel lucky to make strong friendships during the years that I spent at Georgia Tech. Gonca Karakus, Hakan Toreyin, Aytac Atac, Burc Tosyali, Deniz Cinalioglu, Sahin Tetik, Ozge Karakas, Pinar-Gokhan Samur, Bilge-Gokce Gurun, Emrah Uyar, Banu-Evren Ozkaya, Ilgin-Aykagan Ak are a few to name here. I would like to thank Mazlum Kosma and Talia Kayali for their friendship and support. I also would like to thank Pinar Dumlu for being my friend, my sister and supporting me from thousands of miles away.

I am forever indebted to my mom, dad and brother for always being there when I needed them most, and encouraging me when I doubted myself. I would have never made this far without their unconditional love and support. They are the best family in the world.

My heartfelt gratefulness goes to my dear husband Ozan for his love, patience, continuous moral support, understanding, friendship and everything. You even made the most difficult times a good memory in this journey. You always have been on my side and kept me going. You are my true love...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this thesis, we study three applications of dynamic resource allocation: the first two consider dynamic lead-time quotation in make-to-order (MTO) systems with substitutable products and order cancellations, respectively; and the third application is a manpower allocation problem with job-teaming constraints.

Matching supply and demand for manufacturing and service industries has been a fundamental focus of operations management literature, which concentrated on optimizing or improving supply-side decisions since demand has generally been assumed to be exogenously determined. However, recent business trends and advances in consumer behavior modeling have shown that demand for goods and services can clearly be shaped by various decisions that a firm makes, such as price and lead-time. In fact, competition between companies is no longer mainly based on price or product features; lead-time is one of the strategic measures to evaluate suppliers. In MTO manufacturing or service environments that aim to satisfy the customers' unique needs, lead-time quotation impacts the actual demand of the products and the overall profitability of the firm. In the first two parts of the thesis, we study the dynamic lead-time quotation problem in pure MTO (or service) systems characterized by lead-time sensitive Poisson demand and exponentially distributed service times. We formulate the problem as an infinite horizon Markov decision process (MDP) with the objective of maximizing the long-run expected average profit per unit time, where profits are defined to specifically account for delays in delivery of the customer orders.

We study dynamic lead-time quotation problem in two particular settings; one setting with the possibility of demand substitution and another setting with order cancellations. The fundamental trade-off in lead-time quotation is between quoting

short lead-times and attaining them. In case of demand substitution, i.e., in presence of substitutable products and multiple customer classes with different requirements and margins, this trade-off also includes capacity allocation and order acceptance decisions. In particular, one needs to decide whether to allocate capacity to a low-margin order now, or whether to reserve capacity for potential future arrivals of high-margin orders by considering customer preferences, the current workload in the system, and the future arrivals. In the case of order cancellations, one needs to take into account the probability of cancellation of orders currently in the system and quote lead-times accordingly; otherwise quotation of a longer lead-time may result in the loss of customer order, lower utilization of resources, and, in turn, reduced in profits.

In Chapter 2, we study a dynamic lead-time quotation problem in a MTO system with two (partially) substitutable products and two classes of customers. Customers decide to place an order on one of the products or not to place an order, based on the quoted lead-times. We analyze the optimal profit and the structure of the optimal lead-time policy. We also compare the lead-time quotes and profits for different quotation strategies (static vs. dynamic) with or without substitution. Numerical results show that substitution and dynamic quotation have synergetic effects, and higher benefits can be obtained by dynamic quotation and/or substitution when difference in product revenues or arrival rates, or total traffic intensity are higher.

In Chapter 3, we study a dynamic lead-time quotation problem in a MTO system with single product considering the order cancellations. The order cancellations can take place during the period that the order is being processed (either waiting or undergoing processing), or after the processing is completed, at the delivery to the customer. We analyze the behavior of optimal profit in terms of cancellation parameters. We show that the optimal profit does not necessarily decrease as cancellation rate increases through a numerical study. When the profit from a cancelled order, arrival rate of customers, or lead-time sensitivity of customers are high, there is a higher probability that optimal profit increases as cancellation rate increases. We also compare the cancellation scenarios with the corresponding no-cancellation scenarios, and show that there exists a cancellation scenario that is at least as good in terms of profit than a no-cancellation scenario for most of the parameter settings.

In Chapter 4, we study the Manpower Allocation Problem with Job-Teaming Constraints with the objective of minimizing the total completion time of all tasks. The problem arises in various contexts where tasks require cooperation between workers: a team of individuals with varied expertise required in different locations in a business environment, surgeries requiring different composition of doctors and nurses in a hospital, a combination of technicians with individual skills needed in a service company. A set of tasks at random locations require a set of capabilities to be accomplished, and workers have unique capabilities that are required by several tasks. Tasks require synchronization of workers to be accomplished, hence workers arriving early at a task have to wait for other required workers to arrive in order to start processing. We present a mixed integer programming formulation, strengthen it by adding cuts and propose heuristic approaches. Experimental results are reported for low and high coordination levels, i.e., number of workers that are required to work simultaneously on a given task.

CHAPTER I

INTRODUCTION

A fundamental problem addressed through operations research and mathematical modeling is the strategic allocation of limited resources to a set of jobs or tasks. The setting (static, dynamic or stochastic), the properties of the tasks and the resources, and the objective function affect the nature of the problem; but the aim is the same, i.e., allocating the resources considering the characteristics of the resources and the tasks so that the objective is minimized (or maximized). This thesis focuses on dynamic resource allocation problem.

The first two applications are dynamic lead-time quotation problem in make-to-order (MTO) systems, motivated by recent business trends and advances in consumer behavior modeling which show that demand for goods and services are shaped by price and lead-time decisions. With globalization, companies are competing not only on product features and price but also delivering the right product to the right place at the right time (Eskew [19]). Gunasekaran et al. [25] state that product choice is no longer mainly based on price competition; and lead-time is one of the strategic measures to evaluate suppliers. In a 2001 survey, original-equipment manufacturers rated "ability to meet delivery schedules" as the most important factor in selecting a manufacturer, where price ranked the fifth (Ansberry and Aeppel [1]). Quoting a lead-time, defined as the difference between the promised due date of an order and its arrival time, is equivalent to quoting a due date. Quoting reliable lead-times to match supply and demand is a predominant factor since it affects the buyer's production and inventory costs and customer service. For example, the costs of delays are enormous in the semiconductor industry, since production lines need to be shut down if chips are not available (Li [33]). In the automative industry, Saturn fines suppliers \$500 per minute production line stoppages (Frame [20]), and Chrysler fines \$32,000 per hour for a late order (Russell and Taylor [47]). In the aircraft industry, lateness penalties range from \$10,000 to \$1,000,000 per day (George [23], Slotnick and Sobel [49]). The result of wrong lead-time decisions is not only the loss of money but also the loss of goodwill and dissatisfied customers. While companies try to quote short lead-times to attract customers, they are also expected to attain these lead-times not to incur late delivery penalties and loss of goodwill.

In MTO manufacturing or service environments that aim to satisfy the customers' unique needs, lead-time quotation impacts the customers' decisions on whether or not to place an order, the congestion in the system, and the overall profitability of the firm using dynamic lead-time quotation. Depending on its current workload, capacity availability, and anticipated future demand, a firm may offer alternative products or services with similar attributes but different lead times. For example, a shipping company may offer two alternatives to a customer that wants to ship a container from origin port A to destination port B: a direct route with an estimated delivery date in three weeks (due to tight capacity), and a connected route (with stops in ports B and C) with an estimated delivery date in two weeks. While the customer may prefer the direct route, he may choose the connected route due to its earlier delivery date.

In the first application, we examine dynamic lead-time quotation for substitutable products. Dynamic quotation of lead-times considering the customer choice model, and the profits of substitutable products enables a better allocation of resources in order to generate a higher profit. The manufacturer (or service provider) may offer a substitute product (or service) to a customer when the product that customer requests has a long queue which will either result in loss of customer by quoting a long lead-time or lateness penalty (or loss of goodwill) if quoted a short lead-time. Manufacturer may also benefit from the substitute product by planning for future customer arrivals and even directing customer to the product with higher profit. While other researchers modeled dynamic lead-time quotation problem and demand substitution in different settings, to the best of our knowledge no one has specifically studied the dynamic lead-time quotation problem with demand substitution. We analyze the structure of optimal profit and optimal lead-time policies. We then compare the dynamic quotation of substitute products with scenarios where either one or both of products are quoted static and/or substitution is not possible. Furthermore, we investigate the impact of problem parameters (revenues of products, arrival rates of customers, total traffic intensity) and customer preferences (lead-time sensitivity and choice parameters) on benefits obtained by substitution and/or dynamic quotation.

The impact of cancellations can be significant for firms selling expensive customized capital goods such as production equipment, commercial aircraft, medical devices, or defense systems. For example, Qantas Airways recently cancelled the orders of 35 Boeing 787 Dreamliners, which is stated as the largest cancellation for the Dreamliner. Boeing invested an estimated $14 billion or more to develop Dreamliners and these cancellations cloud the efforts to make it profitable (Ostrower [41]). When customers require product delivery within an aggressive lead-time and customized nature of the product makes it risky for suppliers to keep materials in inventory, leading to lengthy and stochastic lead-times; the impact of order cancellations become significant. Order cancellations are observed often in these environments due to rapidly changing technology, intense competition, short product life cycle, and uncertain end product demands.

The second application is dynamic lead-time quotation in a MTO environment considering the possibility of order cancellations. Lead-times quoted ignoring the cancellations may result in under-utilization of resources, and therefore may lead to lower than possible profits. The manufacturer quotes lead-times not only according to current system state information but also cancellation probabilities and related profits. We assume that cancellations may occur after the order is processed as well as while waiting in the queue or processing. While there are other studies modeling the order cancellations, to our knowledge no one has specifically examined the effect of order cancellations on lead-time policy. We derive expressions for time-in-system, expected tardiness duration, and cancellation probability, and then analyze the structure of the optimal profit. We also analyze the impact of cancellation parameters (profit/cost of cancellation and cancellation probability) on the optimal profit and the optimal lead-time quotes. Using a numerical study, we investigate the impact of cancellation rate on the optimal profit under different parameter settings. Furthermore, we compare cancellation scenarios with corresponding no-cancellation scenarios.

The last application is motivated by single-task multi-robot instantaneous assignment problems with heterogeneous tasks and robots. Each task requires a subset of robots to be present simultaneously in order to be accomplished. The problem arises in various contexts where tasks require cooperation between workers. These tasks do not have precedence constraints, i.e., they can be performed in any order; however, they require cooperation between workers which will create a dependency between tasks requiring same workers. For example, in a business environment a team composed of individuals with varied expertise may be required in different locations, each location requiring a different composition of individuals. Consider a hospital where each surgery requires a different composition of doctors and nurses, or service company where a combination of technicians with individual skills is needed to accomplish each task. Therefore, we investigated the problem in general as a manpower allocation problem, i.e., scheduling tasks so that all tasks are completed as earliest as possible. Scheduling tasks means allocating all resources that are required to perform that task for a particular time. If all resources cannot arrive at the task at the same time, then early arriving resources wait until the remaining required resources arrive. We actually determine allocation of resources to tasks in particular times so that total travel time and idle time (while waiting for other required resources) is minimized. The simultaneous presence of workers for task accomplishment combined with the objective of minimizing the total completion time of all tasks is the novel part of our problem. We propose a mixed integer programming (MIP) formulation and strengthen it by cuts. We also propose several heuristic approaches to solve the problem. We compare the heuristics and MIP solutions in order to identify the best way in each setting (in number of tasks and workers and number of required workers by tasks) by extensive numerical tests on randomly generated instances.

CHAPTER II

DYNAMIC LEAD-TIME QUOTATION OF SUBSTITUTABLE PRODUCTS

2.1 Introduction

In make-to-order (MTO) manufacturing or service environments that aim to satisfy customers' unique and customized needs, lead-time quotation significantly impacts customers' decisions on whether or not to place an order. Consequently, lead time decisions can be used to manage the congestion in the system, and, improve the overall profitability of the firm. Quoting short and reliable lead-times is an important competitive advantage among suppliers in many industries. Gunasekaran et al. [25] state that product choice is no longer mainly based on price competition, and is significantly impacted by the lead-time. In fact, lead time is considered to be one of the strategic measures to evaluate suppliers. In a 2001 survey, original-equipment manufacturers rated "ability to meet delivery schedules" as the most important factor in selecting a manufacturer, where price ranked the fifth (Ansberry and Aeppel [1]). While companies try to quote short lead-times to attract customers, they are also faced with the task of meeting these lead-times so as to not incur late delivery penalties and loss of goodwill. The inability to meet quoted due dates can lead to significant penalties for suppliers. For example, the costs of delays are particularly high in the semiconductor industry, since shortages in various components may lead to production line shut-downs (Li [33]). In the automotive industry, Saturn fines suppliers \$500 per minute of production line stoppages (Frame [20]), and Chrysler fines $\$32,000$ per hour for a late order (Russell and Taylor [47]). In the aircraft industry, lateness penalties range from $\$10,000$ to $\$1,000,000$ per day (George [23], Slotnick and Sobel [49]).

Unmet due date quotes not only cause loss of money but also loss of goodwill and erosion of customer service.

Customers have widely varying demands in terms of price and lead-time. Plambeck [44] states that, manufacturers are selling the same product to different customers at different prices based on lead-time in the automotive, electronics and consumer goods industries. For example, corporate fleet buyers may prefer to wait for a long time for a small price discount in the automotive industry. Depending on its current workload, capacity availability, and anticipated future demand; a firm may offer alternative products or services with similar attributes but different lead times. For example, a shipping company may offer two alternatives to a customer who wants to ship a container from origin port A to destination port B: a direct route with an estimated delivery date in three weeks (due to tight capacity), and a connected route (with stops in ports B and C) with an estimated delivery date in two weeks. While the customer may prefer the direct route, he may choose the connected route due to its earlier delivery date.

Dynamic lead time quotation and the availability of (partially) substitutable products helps the manufacturer in matching demand and supply. The manufacturers' trade-off in lead-time quotation is between quoting short lead-times and attaining them. When there are substitutable products, the manufacturer has the flexibility of influencing the demands of the substitutable products; considering the customer preferences, current workload in the system, and the future arrivals. The manufacturer can direct the customer to the item with shorter queue by quoting a shorter lead-time for that item, instead of quoting a short lead-time for the item with a long queue, which may result in lateness penalties. Manufacturer may also benefit from the option of substituting products by planning for future customer arrivals and even directing customer to the product with higher profit. However, this requires the simultaneous consideration of lead-time quotation for both products, significantly increasing the complexity of these decisions.

The manufacturer may shape the demand by quoting lead-times for substitutable products. Although the manufacturing capacities of products may be dedicated, the manufacturer has flexibility at the end-item level through altering the customer choice by quoting lead-times. Demand can be shaped by both price and lead-time decisions, in this thesis we focus on lead-time differentiation, which is the key to make-to-order car manufacturing (Economist [17]). We assume that price and other attributes of products determine the allocation of demand among products before lead-times are quoted. While other researchers modeled dynamic lead-time quotation problem and demand substitution in different settings, to the best of our knowledge no one has specifically studied dynamic lead-time quotation problem with demand substitution. We formulate the problem for two substitutable products, and two customer types with linear choice models. We analyze the structure of optimal profit and optimal lead-time policies. The optimal profit is concave and linear in regions specified by choice parameters and the optimal lead-time policy is not simply monotonic. Then, we evaluate the value of substitution and dynamic quotation through a numerical study. We observe that dynamic quotation and substitution has synergistic effects. Furthermore, we investigate the impact of various problem parameters (i.e., revenues of products, arrival rates of customers, total traffic intensity) and customer preferences (i.e., lead-time sensitivity and choice parameters) on benefits obtained by substitution and/or dynamic quotation. We observe that higher differences in revenues or arrival rates, higher traffic intensity, lead-time sensitivity or cross lead-time elasticity results in higher profit improvements by dynamic quotation and substitution.

This chapter is organized as follows. In Section 2.2, we review the related literature. In Section 2.3, we model and analyze the dynamic lead-time quotation problem for two substitutable products. We assess the benefits of dynamic quotation and substitution under different parameter settings in Section 2.5. Summary of major findings and conclusions are presented in Section 2.6.

2.2 Literature Review

The related literature can be classified into two categories: lead-time quotation, and demand substitution. For detailed reviews of lead-time management literature, readers are referred to Kaminsky and Hochbaum [29] and Keskinocak and Tayur [32]. In most of the previous literature on due-date management, an implicit assumption is that each customer who is quoted a due date places an order, i.e., order acceptance decisions are ignored. However, quoted due dates may have a strong affect on order placement decisions. Charnsirisakskul et al. [8, 9], Kapuscinksi and Tayur [30] and Keskinocak et al. [31] consider order acceptance decisions assuming deterministic processing times, whereas Ata [2] and Ata and Olsen [3] model admission control and sequencing decisions of a system manager in a make-to-order production system, making heavy traffic assumptions. Duenyas [15] and Duenyas and Hopp [16] also consider order acceptance decisions and they use stochastic processing times. Duenyas and Hopp [16] study a lead-time quotation and order-sequencing problem in an MTO environment with lead-time sensitive customers. They show that the optimal due-date quotes are increasing in the number of orders in the queue, assuming exponential interarrival and service times in the presence of first-come first-served (FCFS) processing. Furthermore, assuming the firm may choose a scheduling policy other than FCFS, Duenyas and Hopp [16] show that the optimal due-date quotation/order sequencing policy is the earliest due-date sequence. Duenyas [15] extends this study to two customer classes and proposes a heuristic that performs well relative to other due-date setting rules in the literature. Recently, Savasaneril et al. [48] extend the dynamic lead-time quotation problem to hybrid make-to-order/make-to-stock environments, i.e., a base-stock inventory system characterized by lead-time sensitive demand. They show that the lead-time quotes are lower in an MTO system compared to a base-stock system.

Our work is most closely related to Duenyas and Hopp [16]. We study the dynamic lead-time quotation problem in an MTO environment where lead-time sensitive customers arrive to a system of two substitutable products, each processed in a FCFS order. Since we consider two substitutable products, we analyze the behavior of lead-time quotes in terms of two state variables, namely the number of orders in the system for both products. Furthermore, we compare the profits of scenarios with and without substitution and/or dynamic versus static lead-time quotation under different parameter settings.

Most of the research in demand substitution has focused on assortment and inventory decisions in a make-to-stock (MTS) environment. Gaur and Honhon [21], Smith and Agrawal [50], and van Ryzin and Mahajan [53] consider the problem of assortment and inventory decisions using multinomial logit (MNL), general demand, and locational choice models, respectively. Both van Ryzin and Mahajan [53] and Smith and Agrawal [50] assume static substitution, i.e., a customer's choice is not affected by the current inventory levels, and if a customer selects a variant that is out of stock, the sale is lost. On the other hand, customer chooses among the available ones after observing the current inventory levels in dynamic substitution. Gaur and Honhon [21] consider both static and dynamic substitution. They determine optimal assortment and inventory policies for static substitution and propose heuristics under dynamic substitution. Mahajan and Ryzin [39] state that static substitution is a somewhat unsatisfying assumption and investigates the effect of dynamic substitution on assortment planning. Our model also assumes dynamic substitution, i.e., lead-times are quoted according to the current workload in the system affecting the customer choice. Maddah and Bish [38] consider pricing decisions in addition to assortment and inventory decisions using a MNL choice model in a newsvendor type inventory setting. They propose a dominance relationship that simplifies the search for an optimal assortment. Assortments having items with the largest average margins are not necessarily the optimal ones but yield expected profits close to the optimal profits. Hence, they propose a simple heuristic based on equal margins.

Pricing decisions for substitutable products over a predetermined, finite selling season are considered by Bitran et al. [5], Dong et al. [14] and Suh and Aydin [52] in a retail setting. Bitran et al. [5] study optimal pricing policies for a family of substitutable perishable products with demand correlation and propose a pricing policy that maximizes expected cumulative revenues over a finite selling horizon. They propose a Walrasian Choice model (WAL) which establishes a single (absolute) ranking of the products based on non-price attributes to incorporate customer segmentation, and derive a price-sensitive demand function capturing the buyers' purchasing behavior. Dong et al. [14] study both the dynamic pricing and initial inventory decisions for a retailer who faces a long supply lead-time and a short selling season. They compare static, unified dynamic (identical price to all variates and this price is changed dynamically) and mixed dynamic pricing (combination of static and dynamic pricing) strategies with full-scale dynamic pricing and find that full-scale dynamic pricing improves profits significantly in presence of inventory scarcity. Furthermore, the initial inventory levels are found to be relatively insensitive to the pricing scheme employed and they propose a computationally efficient approach for the initial inventory decision. Suh and Aydin [52] consider dynamic pricing of two substitutable products over a predetermined selling season where initial inventory levels are fixed exogenously. They show that the marginal value of an item increases in the remaining time and decreases in its stock level and the other product's stock level. However, the optimal price is not simply monotonic in the remaining time or the other product's stock level. In our work, instead of inventory and pricing decisions in a retail setting, we consider lead-time decisions in a MTO setting.

Ceryan et al. [7] consider joint price- and capacity-based substitution to alleviate the level of inventory mismatches of substitutable products. They consider a firm producing two products using capacitated product-dedicated and flexible resources. They show that stable price differences can be maintained across items with an available flexible resource, which may establish consistent price positioning even under a dynamic pricing strategy. In our work, we focus on the flexibility achieved by demand substitution and dynamic lead-time quotation.

In our work, instead of assortment, pricing or inventory decisions, we study the relationship between the strategies of demand substitution and dynamic lead-time quotation. We analyze the structure of optimal profit and the optimal lead-time quotation policy. We also evaluate the value of dynamic quotation and/or substitution by comparing the profit improvements obtained by switching to dynamic quotation from static quotation and/or addition of substitution under different parameter settings.

2.3 Model Formulation

We model a make-to-order (MTO) manufacturer who offers two substitutable products that are manufactured on separate dedicated lines, indexed by $i = 1, 2$. Service times of products are independent exponential random variables with mean μ_i. We assume that each product enjoys a market share in the sense that a certain fraction of the total customer base of the company prefers product 1 (2), even though would entertain the idea of buying product 2 (1). Customers with a preference for product 1 (2) are labeled as customer type 1 (2). We assume that customer orders arrive according to a Poisson process with rate λ_c for customer type $c = 1, 2$. When a customer arrives, the customer is provided lead-time quotes d_1 and d_2 for product 1 and product 2, respectively. For each product 1 and 2, $d_i \in (0, d_i^{\max}]$, and we assume that the lead time quotes are strictly positive, i.e., $d_i > 0$ due to MTO setting. Upon receiving the lead time quotes for each product, the customer either places an order for one of the products or leaves the system without placing an order. Below, we will discuss our assumptions on the probabilities of these events. If a customer places an order, he/she has to wait until the order is processed and becomes available. If customer's order is completed after the quoted lead-time, a tardiness penalty is incurred; the tardiness penalty increases linearly with the tardiness duration at rate l_i for product i. The objective is to maximize the long-run average expected profit per unit time, where profit from a customer is equal to the revenue (r_i for product i) minus late delivery penalties incurred for that order.

The probabilities of the customer's decisions on whether to place an order, and for which product, are described by a linear choice model. We assume that a customer of type c puts in an order for product i is given by , $p_{ci}(d_1, d_2)$, which is given in Equations (1-2) for $c = i$, $i = 1, 2$ and $j = 3 - i$.

$$p_{ii}(d_1, d_2) = \left[\frac{1}{d_i^{\max}}(d_i^{\max} - d_i) - \frac{k_{ij}}{d_j^{\max}}(d_j^{\max} - d_j)\right]^+, \tag{1}$$

$$p_{ij}(d_1, d_2) = \left[-\frac{k_{ii}}{d_i^{\max}}(d_i^{\max} - d_i) + \frac{k_{ii} + k_{ij}}{d_j^{\max}}(d_j^{\max} - d_j)\right]^+, \tag{2}$$

where $(.)^+$ denotes $\max(., 0)$.

In our model, cross lead-time elasticity for choice probability of customer c for product i is denoted by $\frac{k_{ci}}{d_i^{\max}}$. For instance, everything else being equal, one unit increase in the lead-time of product 2 results in an increase of $\frac{k_{12}}{d_2^{\max}}$ in the choice probability of customer 1 for product 1. We make the following assumptions:

A1. All parameters are positive: $k_{ij} > 0$.

The choice probability of product 1 (2) is non-increasing in d_1 (d_2), and non-decreasing in d_2 (d_2).

A2. $k_{12} < 0.5$ and $k_{21} < 0.5$.

Customer type 1 chooses product 1, and customer type 2 chooses product 2 when they are quoted the minimum lead-times for both products, i.e., $p_{11}(\epsilon, \epsilon) > p_{12}(\epsilon, \epsilon)$ and $p_{22}(\epsilon, \epsilon) > p_{21}(\epsilon, \epsilon)$.

A3. $k_{11} + k_{12} \leq 1$ and $k_{21} + k_{22} \leq 1$.

A unit increase in the lead-time of a product creates a larger decrease in its own choice probability than the increase in the other product's choice probability for a given customer type. The choice probability of a customer is decreasing in the lead-times of both products by this assumption.

A4. $k_{12} < \frac{d_2^{\max}}{d_1^{\max}}$ and $k_{21} < \frac{d_1^{\max}}{d_2^{\max}}$.

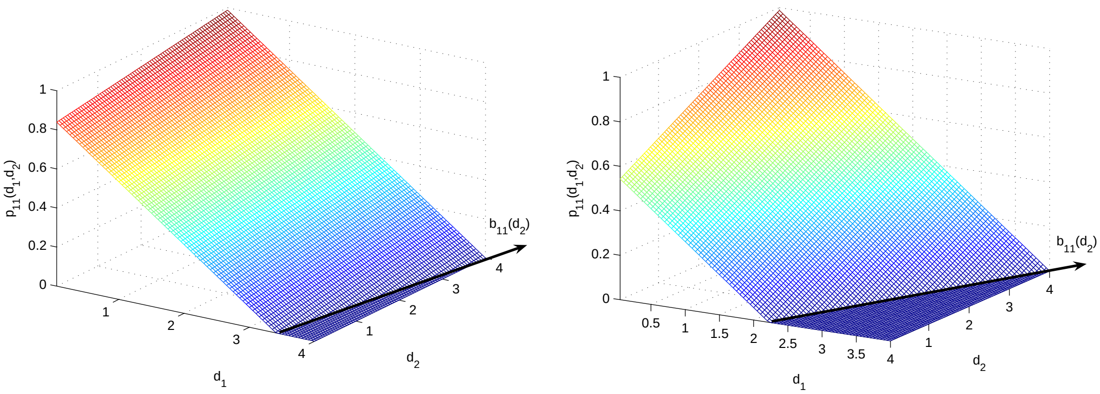
The choice probability of the preferred product is affected more by a unit change in its own lead-time than a unit change in the other product's lead-time.

A5. When the maximum lead-times are quoted for both products, the customer will leave without putting in an order with probability one, i.e., $p_{c0}(d1, d2) = 1$.

Note that the choice model is linear in the quoted lead-time, and cross lead-time effects are similar to those considered in Pekgun et al. [43] and Ceryan et al. [7]. Pekgun et al. [43] uses a demand model which is linear in quoted price and lead-time, and cross price and cross lead-time effects. Ceryan et al. [7] defines the demand model in a linear additive fashion using individual and cross-price elasticities for each product type. Since we focus only on lead-time decisions, our demand model is also similar to the one used in Savasaneril et al. [48] where they use convex and concave as well as linear models to define acceptance probability.

The impact of parameters k_{11} and k_{12} on choice probability $p_{11}(d_1, d_2)$ are shown in Figure 1. As the cross lead-time elasticity for product 2, i.e., k_{12}, increases, the effect of d_2 on the choice probability of product 1 for customer 1, i.e., $p_{11}(d_1, d_2)$, increases.

Note that for all values of c and i, and $j = 3-i$, the choice probability $p_{ci}(d_1, d_2) =$



(a) $k_{12} = 0.15$ (b) $k_{12} = 0.45$

Figure 1: Impact of increasing k_{12} on $p_{11}(d_1, d_2)$ where $d_1^{\max} = d_2^{\max} = 4$

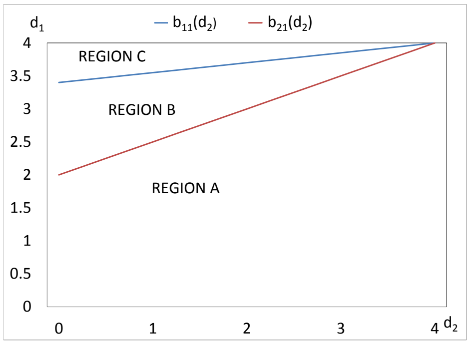
0 for all $d_i \geq b_{ci}(d_j)$, where the boundary points, $b_{ci}(d_j)$, are defined as follows:

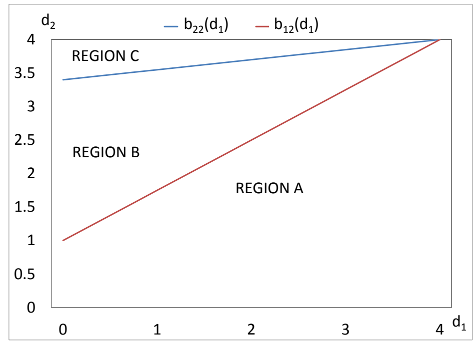
$$b_{ii}(d_j) = d_i^{\max}\left(-\frac{k_{ij}}{d_j^{\max}}(d_j^{\max} - d_j) + 1\right), \tag{3}$$

$$b_{ij}(d_i) = \frac{d_j^{\max}}{k_{ii} + k_{ij}}\left(-\frac{k_{ii}}{d_i^{\max}}(d_i^{\max} - d_i) + k_{ii} + k_{ij}\right). \tag{4}$$

The boundary points of choice probabilities, $b_{ci}(d_j)$ (as defined in Equations (3) and (4)) and the boundaries of the action space, namely zero and $d^{\max}$ values, define three regions for choice probabilities: (i) Region A: the choice probabilities of the product for both customers are positive, i.e., between 0 and the minimum of boundary points, (ii) Region B: the choice probability of the product for only one of the customers is positive, i.e., between two boundary points, (iii) Region C: the choice probability of the product for both customers are 0 at the given level of other product's lead-time quote, i.e., between the maximum of boundary points and the associated $d^{\max}$ value. The size of the regions depend on choice probability parameters as shown in Figure 2 for $d_1^{\max} = d_2^{\max} = 4$, and choice probability parameters $k_{11} = 0.45$ and $k_{12} = k_{21} = k_{22} = 0.15$. Region A increases and, Regions B and C decrease as the other product's lead-time quote increases since other product becomes less likely to be chosen with higher lead-time quote.

The problem is formulated as an infinite-horizon Markov decision process (MDP)





(a) For choice of product 1, $p_{11}(d_1, d_2)$ and $p_{21}(d_1, d_2)$

(b) For choice of product 2, $p_{22}(d_1, d_2)$ and $p_{12}(d_1, d_2)$

Figure 2: Regions A, B and C according to boundary points

model with the long-run average expected profit per unit time criteria. The manufacturer quotes lead-times for each product each time a customer arrives to the system, where the lead time quote for product i, $d_i \in (0, d_i^{\max}]$. The state of the system at time t is denoted by the vector, $x(t) = (x_1(t), x_2(t))$, where $x_i(t)$ denotes the number of product i orders in the system at time t. We drop t from the notation, and simply use x_i to denote the state. The continuous-time model is converted to an equivalent discrete-time model through uniformization (Lippman [36]), and the corresponding optimality equation is given in Equation (5), where g^* denotes the optimal long-run expected average profit per unit time, and $v^*(x_1, x_2)$ is the relative value of starting in state (x_1, x_2) under the optimal policy.

$$\begin{aligned}\frac{g^*}{\nu} + v^*(x_1, x_2) = \max_{d_1 \in (0, d_1^{\max}], d_2 \in (0, d_2^{\max}]} \Bigg\{ \bigg(\frac{\mu_1}{\nu} v^*((x_1-1)^+, x_2) + \frac{\mu_2}{\nu} v^*(x_1, (x_2-1)^+) \\ + \frac{\lambda_1}{\nu} p_{11}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2) \bigg) (r_1 - l_1 L(x_1, d_1) + v^*(x_1+1, x_2)) \\ + \bigg(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2) \bigg) (r_2 - l_2 L(x_2, d_2) + v^*(x_1, x_2+1)) \\ + \bigg(\frac{\lambda_1}{\nu} p_{10}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{20}(d_1, d_2) \bigg) v^*(x_1, x_2) \Bigg\}, \quad (5)\end{aligned}$$

where $\nu = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$ is the uniformization rate, and $L(x_i, d_i)$ is the expected delay when a customer who places an order for product i arrives in state x_i and is quoted lead-time d_i. Assuming the system operates under the FCFS discipline for each product line, $L(x_i, d_i)$ can be calculated as

$$L(x_i, d_i) = \int_{d_i}^{\infty} (y - d_i) h_{x_i+1}(y) dy \ , \tag{6}$$

where $h_{x_i}(.)$ denotes the probability density function of Erlang distribution with parameters μ_i and x_i and is equal to

$$h_{x_i}(y) = \frac{\mu_i(\mu_i y)^{x_i - 1}}{(x_i - 1)!} e^{-\mu_i y} \tag{7}$$

The first two terms in Equation (5) refer to the cases where the customer accepts the quote and places an order for product 1 or 2, and the state changes to (x_1+1, x_2) or (x_1, x_2+1), respectively. The third term refers to the case where the customer leaves the system without placing an order. The fourth and fifth terms refer to the cases when the processing of products 1 and 2, respectively, are completed.

Lemma 1 *$L(x_i, d_i)$ is convex and decreasing in $d_i \geq 0$, $i = 1, 2$.*

Proof. Proof. The proofs of all propositions, lemmas and theorems are included in Appendix A.4. ■

In Lemma 1, we prove that the expected tardiness duration is convex and decreasing in lead-time quotes. We use the properties of the expected tardiness duration in proving properties of relative value functions, and optimal policy.

2.4 Optimal Policy Analysis

In this section, we characterize the optimal solution to (5). The optimal solution at state (x_1, x_2) is denoted by $(d_1^*(x_1, x_2), d_2^*(x_1, x_2))$ or simply (d_1^*, d_2^*). Let us define the burden brought on by one additional product 1 and product 2 order under optimal policy as $\triangle_1 v^*(x_1, x_2) = v^*(x_1, x_2) - v^*(x_1+1, x_2)$ and $\triangle_2 v^*(x_1, x_2) = v^*(x_1, x_2) - v^*(x_1, x_2+1)$, respectively. Using the $\triangle$ notation and replacing $p_{c0}(d_1, d_2)$ by $(1 - p_{c1}(d_1, d_2) - p_{c2}(d_1, d_2))$, Equation (5) reduces to the following:

$$
\begin{aligned}
\frac{g^*}{\nu} + v^*(x_1, x_2) &= \max_{d_1 \in (0, d_1^{\max}], d_2 \in (0, d_2^{\max}]} \left\{ \frac{\lambda_1 + \lambda_2}{\nu} v^*(x_1, x_2) \right. \\
&+ \left(\frac{\lambda_1}{\nu} p_{11}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2) \right) (r_1 - l_1 L(x_1, d_1) - \triangle_1 v^*(x_1, x_2)) \\
&+ \left(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2) \right) (r_2 - l_2 L(x_2, d_2) - \triangle_2 v^*(x_1, x_2)) \\
&+ \left. \frac{\mu_1}{\nu} v^*((x_1 - 1)^+, x_2) + \frac{\mu_2}{\nu} v^*(x_1, (x_2 - 1)^+) \right\}. \qquad (8)
\end{aligned}
$$

According to Equation (8), (d_1^*, d_2^*) maximizes

$$
\begin{aligned}
\Pi^*_{x_1,x_2}(d_1, d_2) = &\left(\frac{\lambda_1}{\nu} p_{11}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2) \right) (r_1 - l_1 L(x_1, d_1) - \triangle_1 v^*(x_1, x_2)) \\
&+ \left(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2) \right) (r_2 - l_2 L(x_2, d_2) - \triangle_2 v^*(x_1, x_2)). \qquad (9)
\end{aligned}
$$

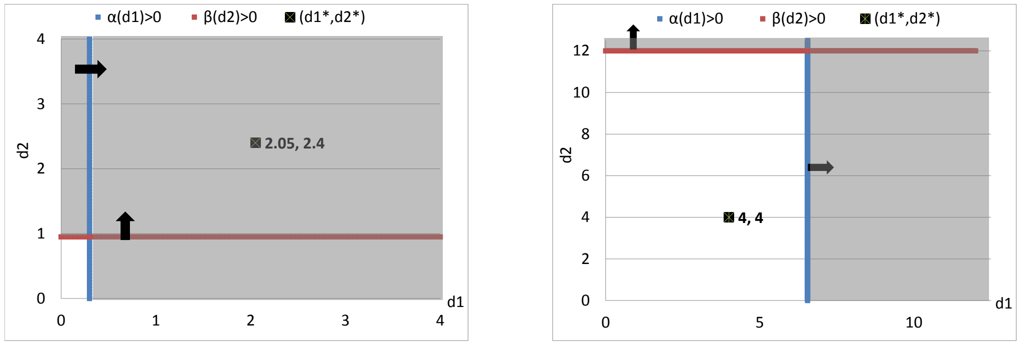
Below we show that (d_1^*, d_2^*) at any state (x_1, x_2) is confined within the region defined as follows:

$$
\Gamma_{x_1,x_2} = \left\{ d_1 \in (0, d_1^{\max}], d_2 \in (0, d_2^{\max}] : \alpha_{x_1,x_2}(d_1) \geq 0 \text{ or } \beta_{x_1,x_2}(d_2) \geq 0 \right\}, \qquad (10)
$$

where $\alpha_{x_1,x_2}(d_1) = r_1 - l_1 L(x_1, d_1) - \triangle_1 v^*(x_1, x_2)$ and $\beta_{x_1,x_2}(d_2) = r_2 - l_2 L(x_2, d_2) - \triangle_2 v^*(x_1, x_2)$ denote the expected profit (revenue minus tardiness penalty) minus the additional burden of one more order for product 1 and 2, respectively.

Proposition 1 *The optimal lead-time quotes either satisfy $(d_1^*, d_2^*) \in \Gamma_{x_1,x_2}$, or $\Gamma_{x_1,x_2} = \emptyset$ and $(d_1^*, d_2^*) = (d_1^{\max}, d_2^{\max})$.*

Figure 3 depicts the set of optimal lead-time quotes with respect to Γ_{x_1,x_2} calculated by using the optimal relative value functions for the instance in Example 1. At state $(x_1, x_2) = (3, 1)$, $d_1^* \geq 0.3$, or $d_2^* \geq 0.95$ are sufficient for $\alpha_{x_1,x_2}(d_1)$ and $\beta_{x_1,x_2}(d_2)$ to be non-negative, and the optimal lead-time quotes are $(2.05, 2.4)$. On the other hand, at state $(x_1, x_2) = (6, 6)$, $d_1^* \geq 6.55$ and $d_2^* \geq 12$ so that $\alpha_{x_1,x_2}(d_1)$ and $\beta_{x_1,x_2}(d_2)$ are non-negative. However $d_1^{\max} = d_2^{\max} = 4$ and $\Gamma_{x_1,x_2} = 0$. Hence,



(a) $x_1 = 3, x_2 = 1$ (b) $x_1 = 6, x_2 = 6$

Figure 3: Γ_{x_1,x_2} and optimal lead-time quotes for Example 1: $r_1 = 15$, $r_2 = 5$, $\lambda_1 = \lambda_2 = 0.15$, $d_1^{\max} = d_2^{\max} = 4$, $k_{11} = k_{12} = k_{21} = k_{22} = 0.15$, $\mu_1 = \mu_2 = 0.5$ and $l = 1.5$

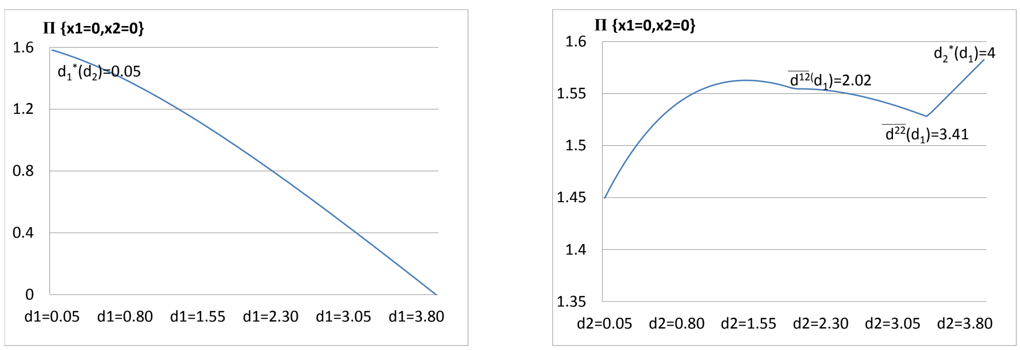
the optimal lead-time quotes are $(4, 4)$ which are the maximum lead-time values corresponding to rejecting the customer at state $(x_1, x_2) = (6, 6)$. $\alpha_{x1=6,x_2=6}(d_1) \geq 0$ corresponds to $d_1^* \geq 6.55$ which means that the expected profit minus the burden brought by one additional product 1 order will be non-negative only if the quoted lead-time is ≥ 6.55. Similarly for product 2, $\beta_{x_1=6,x_2=6}(d_1) \geq 0$ correponds to $d_2^* \geq 12$.

After defining the set of optimal lead-time quotes, we further explore the structure of the optimal policy. In Lemma 2, we show that the relative value functions are non-increasing in state variables and then, using this result we show the behavior of $\Pi^*_{x_1,x_2}(d_1, d_2)$ in terms of d_1 and d_2 and the properties of the optimal policy.

Lemma 2 *The relative value functions under optimal policy, $v^*(x_1, x_2)$, are non-increasing in state variables x_1 and x_2.*

As the number of orders in the system increases for product i, then the expected profit from an order placed for product i decreases (or remains constant) ignoring the tardiness penalty. For single product case, Duenyas and Hopp [16] show that relative value functions are non-increasing in state. Cil [11] show that relative value functions are non-increasing in state variables in dynamic pricing problem of a queueing system with two classes of customers. In dynamic lead-time quotation for substitutable products, lead-time quotes appear in choice probability and tardiness duration compared to Duenyas and Hopp [16] and Cil [11]. We use the properties of tardiness duration and optimality conditions together in order to prove that relative value functions are non-increasing in state variables.

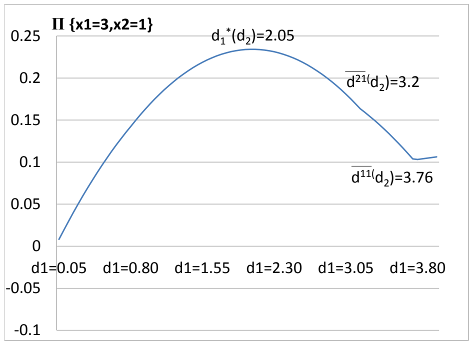
Lemma 3 $\Pi^*_{x_1,x_2}(d_1, d_2)$ *is concave in regions A and B, and linear in region C with respect to* d_1 *(*d_2*) for a given* d_2 *(*d_1*).*
Furthermore, the sign of $\beta_{x_1,x_2}(d_2)$ *(*$\alpha_{x_1,x_2}(d_1)$*) determines whether* $\Pi^*_{x_1,x_2}(d_1, d_2)$ *is increasing or decreasing in region C with respect to* d_1 *(*d_2*).*



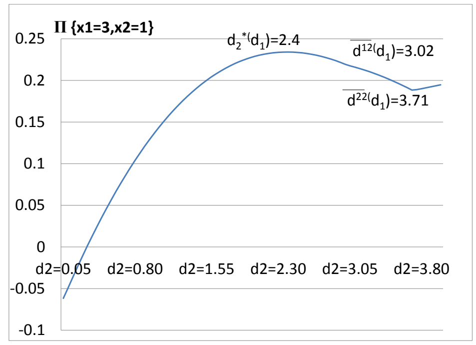
(a) $d_2 = 4$ (b) $d_1 = 0.05$

Figure 4: Behavior of $\Pi^*_{x_1=0,x_2=0}$ w.r.t. d_1 and d_2

Lemma 3 states that for a given d_2 value, the behavior of $\Pi^*_{x_1,x_2}(d_1, d_2)$ with respect to d_1, i.e., the part in the objective that could be changed with actions, depends on whether customer 1 and/or 2 have a positive probability of choosing product 1 as shown in Figures 4 and 5. If at least one of the customers has a positive probability of choosing product 1 (i.e., Regions A and B), $\Pi^*_{x_1,x_2}(d_1, d_2)$ is concave. In these regions, the choice probability of product 1 decreases whereas the expected profit $(r_1 - l_1 L(x_1, d_1))$ increases due to lateness cost decreasing, and the expected profit from the choice of the other product increases as d_1 increases. However, if both



(a) $d_2 = 2.4$



(b) $d_1 = 2.05$

Figure 5: Behavior of $\Pi^*_{x_1=3,x_2=1}$ w.r.t. d_1 and d_2

customers have zero probability of choosing product 1, then the expected profit from the choice of product 1 becomes zero, and the expected profit from the choice of product 2 increases as d_1 increases. Note that the expected profit from the choice of product 2 may be positive or negative.

Let us define $\pi_1(x_1, x_2)$ and $\pi_2(x_1, x_2)$ as the revenue minus the burden brought by an additional order for product 1, $r_1 - \triangle_1 v(x_1, x_2)$, and product 2, $r_2 - \triangle_2 v(x_1, x_2)$, respectively.. If $\pi_1(x_1, x_2)$ is very large compared to $\pi_2(x_1, x_2)$, then it would be optimal to quote the minimum lead-time for product 1. Since in this situation, one does not want to decrease the choice probability of product 1 in return to increasing the choice probability of product 2. However, when $\pi_1(x_1, x_2)$ and $\pi_2(x_1, x_2)$ are close, it may be more profitable to trade-off between the choice probabilities of two products, i.e., decreasing the choice probability of product 1 and increasing the choice probability of product 2 by increasing d_1. The concavity of $\Pi^*_{x_1,x_2}(d_1, d_2)$ in lead-time quotes arises from the lateness cost which decreases as d_1 increases, increasing the profit from the choice of product 1.

First we define the optimality equations which are the first derivatives of $\Pi^*_{x_1,x_2}(d_1, d_2)$

with respect to d_1 and d_2, respectively:

$$\frac{\partial \Pi^*_{x_1,x_2}(d_1,d_2)}{\partial d_1} = \frac{\lambda_1 p_{11}(d_1,d_2)+\lambda_2 p_{21}(d_1,d_2)}{\nu}(-l_1\frac{\partial L(x_1,d_1)}{\partial d_1}) \tag{11}$$

$$-\frac{\lambda_1 \mathbf{1}\{p_{11}(d_1,d_2)>0\}+\lambda_2(k_{21}+k_{22})\mathbf{1}\{p_{21}(d_1,d_2)>0\}}{\nu d_1^{\max}}\alpha_{x_1,x_2}(d_1) \tag{12}$$

$$+\frac{\lambda_1 k_{12}+\lambda_2 k_{22}}{\nu d_1^{\max}}\beta_{x_1,x_2}(d_2) = 0. \tag{13}$$

$$\frac{\partial \Pi^*_{x_1,x_2}(d_1,d_2)}{\partial d_2} = \frac{\lambda_1 p_{12}(d_1,d_2)+\lambda_2 p_{22}(d_1,d_2)}{\nu}(-l_2\frac{\partial L(x_2,d_2)}{\partial d_2}) \tag{14}$$

$$-\frac{\lambda_1(k_{11}+k_{12})\mathbf{1}\{p_{12}(d_1,d_2)>0\}+\lambda_2 \mathbf{1}\{p_{22}(d_1,d_2)>0\}}{\nu d_2^{\max}}\beta_{x_1,x_2}(d_2) \tag{15}$$

$$+\frac{\lambda_1 k_{12}+\lambda_2 k_{22}}{\nu d_2^{\max}}\alpha_{x_1,x_2}(d_1) = 0\ . \tag{16}$$

In Theorem 1, we state the optimality conditions for d_1 given d_2^*; similar arguments hold for d_2 given d_1^*. The optimal pair (d_1^*, d_2^*) satisfies both conditions.

Theorem 1 *The optimal lead-time quote for product* 1, d_1^*, *at any state* (x_1, x_2) *has the following properties:*

(a) If $\alpha_{x_1,x_2}(d_1^{\max}) < 0$, *then* $d_1^* = d_1^{\max}$.

(b) If $\alpha_{x_1,x_2}(d_1^{\max}) \geq 0$*:*

(i) if $\beta_{x_1,x_2}(d_2^{\max}) < 0$ *:*

- *if* $\frac{\partial \Pi^*}{\partial d_1}|_{d_1=\epsilon} < 0$, *then* $d_1^* = \epsilon$.
- *Otherwise,* $\exists\ d_1^*$ *s.t.* $\frac{\partial \Pi^*}{\partial d_1}|_{(d_1^*, d_2^{\max})} = 0$.

(ii) If $\beta_{x_1,x_2}(d_2^{\max}) > 0$, *the set of candidates for* d_1^* *are* $S_1(d_2^*)$*:*

- ϵ *if* $\frac{\partial \Pi^*}{\partial d_1}|_{d_1=\epsilon} < 0$,
- *If* $\exists$ *a* $\hat{d}_1$ *s.t.* $p_{11}(\hat{d}_1, d_2^*) + p_{21}(\hat{d}_1, d_2^*) > 0$ *and* $\frac{\partial \Pi^*}{\partial d_1}|_{d_1=\hat{d}_1} = 0$, *then* $d_1^* = \hat{d}_1$,
- $\hat{d}_1 = min(b_{11}(d_2^*), b_{21}(d_2^*))$ *if* $\frac{\partial \Pi^*}{\partial d_1}|_{d_1=\hat{d}_1-} > 0$ *and* $\frac{\partial \Pi^*}{\partial d_1}|_{d_1=\hat{d}_1+} < 0$,

- $d_1^{\max}$ *if* $\beta_{x_1,x_2}(d_2^*) > 0$.

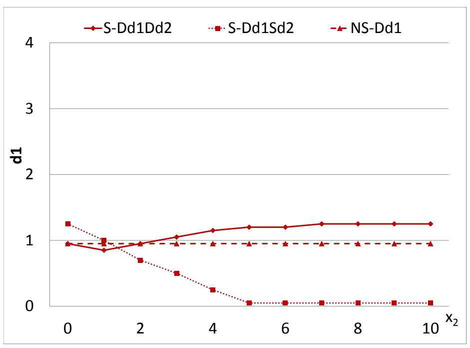
(The $-$ and $+$ next to the points of differentiation indicate the left and right derivative, respectively.)

Theorem 1 defines the set of candidates for d_1^* which consists of one, two or four elements according to the signs of $\alpha_{x_1,x_2}(d_1^{\max})$ and $\beta_{x_1,x_2}(d_2^{\max})$. Calculation of $\alpha_{x_1,x_2}(d_1^{\max})$ and $\beta_{x_1,x_2}(d_2^{\max})$ require the optimal value functions, i.e., $v^*(x_1,x_2)$. Instead of $\alpha_{x_1,x_2}(d_1^{\max})$, we use $r_1 - l_1 L(x_1, d_1^{\max})$. Since if $r_1 - l_1 L(x_1, d_1^{\max}) < 0$, then $\alpha_{x_1,x_2}(d_1^{\max})$ is also negative by Lemma 2. A similar argument holds for $\beta_{x_1,x_2}(d_2^{\max})$.

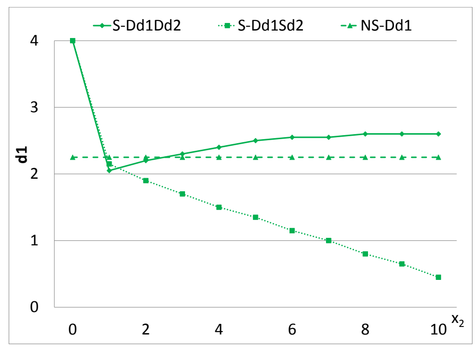
Proposition 2 *If* $r_1 - l_1 L(x_1, d_1^{\max}) < 0$, *then* $d_1^* = d_1^{\max}$.
If $r_2 - l_2 L(x_2, d_2^{\max}) < 0$, *then* $d_2^* = d_2^{\max}$.

Figure 4 depicts the behavior of $\Pi^*_{x_1=0,x_2=0}$ with respect to d_1 and d_2, respectively, for the optimal lead-time quoted $d_2 = 4$ and $d_1 = 0.5$ in Example 1 defined after Proposition 1. $\Pi^*_{x_1,x_2}$ is concave between $(0, 2.02]$ and $[2.02, 3.41]$ and linearly increasing in $[3.41, 4]$ with respect to d_2 for $d_1 = 0.05$. The optimal d_2 is $d_2^{\max} = 4$, which corresponds to the case $(b)-(ii)$ of Theorem 1 for d_1 where one of the four possible values is optimal. For $d_2 = 4$, only region A exists and $\Pi^*_{x_1,x_2}$ is concave and the derivative of $\Pi^*_{x_1,x_2}(d_1,d_2)$ at $d_1 = 0.05$ is negative, i.e., decreasing. Therefore, the optimal d_1 is 0.05. On the other hand, for state $(x_1,x_2) = (3,1)$, $\Pi^*_{x_1,x_2}$ is concave in the first two regions and linearly increasing in Region C with respect to both d_1 and d_2. The optimal lead-times occur at the maximizing point of Region A in both cases.

Duenyas and Hopp [16] show that optimal lead-times are non-decreasing in the number of orders in the system for one product. In Figure 6 we label this case as NS-Dd1 to represent the case where there is only product 1, which is quoted lead times dynamically. We see that the lead time d_1 quoted for $x_1 = 1$ is significantly lower than that for $x_1 = 2$), as also predicted by [16]. In the case with substitutable products, the case with dynamic lead time quotes for both products (labeled as S-Dd1Dd2) and the case with dynamic lead time quote for only one product (labeled as S-Dd1Sd2 below) show different behavior. The optimal d_1 quotes are nonincreasing in x_2 when product 2 is quoted static lead times, whereas the optimal d_1 quotes first decrease and then increase (or remain constant) as x_2 increases. Figure 6 depicts the behavior of optimal policies for different scenarios. Note that the static lead time value used for product 2 below in the S-Dd1Sd2 case is 1.5.



(a) $x_1 = 1$



(b) $x_1 = 2$

Figure 6: Comparison of optimal lead-time quote for product 1, d_1 for scenarios NS-Dd1, S-Dd1Sd2 and S-Dd1Dd2 for Example 3 $r_1 = r_2 = 15, \lambda_1 = \lambda_2 = 0.45, d_1^{\max} = d_2^{\max} = 4, k_{11} = k_{12} = k_{21} = k_{22} = 0.15$.

Observation 1 *When there exists a substitutable product j which is quoted static lead-times, the lead-time quote of the product i remains non-decreasing in number of orders of product i. Whereas in terms of number of orders of product j, it is no longer constant due to substitution effect.*

Observation 1 states that d_i^* is non-increasing in x_j, starting from a higher quote for lower x_j's and then lower quotes compared to the constant level of no-substitution case for higher x_j's. Since d_2 cannot be changed with respect to x_1 or x_2, the only way to lower choice probability of product 2 is decreasing d_1^* as queue gets longer for product 2. Decreasing d_1^* as x_2 increases not only decreases p_{c2} but also increases p_{c1} as product 2 gets longer.

Observation 2 *When there is a substitutable product j which is also quoted dynamically, the lead-time quote of the product i remains non-decreasing in number of orders of product i. Whereas in terms of number of orders of product j it is no longer monotonic due to substitution effect.*

Compared to no-substitution lead-time quote, dynamic substitution lead-time quote may start at a higher value and decrease to a lower value, and then slightly increase to a higher than or equal to value than the values quoted in the case with static lead times quoted for the other product. Also, it should be noted that dynamic substitution quote stays much closer to the no-substitution values whereas static-dynamic values may decrease to much lower values. This observation is similar to the one made by Suh and Aydin, who in the context of a dynamic pricing problem of two substitutable products showed that the optimal price is not monotonic in other product's stock level [52]. They show that the marginal value of an item is decreasing in its own stock level and other product's stock level.

2.5 Performance of Substitution and/or Dynamic Quotation

In this section, we compare Substitution(S) vs. no-substitution (NS) and dynamic (D) vs. static (S) scenarios (see Figure 7) to the base case of no-substitution, static d_1 and static d_2 (NS-Sd1Sd2). In static quotation scenario, we assume that the lead-time quotes are not state dependent, i.e., same for all states. The attractiveness of static quotation is that the lead-time decision does not require real-time state information. We conduct an extensive computational study investigating the impact of the revenue difference between the products (i.e., $r_2 - r_1$), arrival rates (i.e., $\frac{\lambda_1}{\lambda_1+\lambda_2}$), total traffic intensity (i.e., $\lambda_1+\lambda_2$), lead-time sensitivity ($d_2^{\max}$) and cross lead-time elasticity (k_{12})

on substitution and/or dynamic quotation. We highlight the important insights in this section.

We investigate the following topics:

- Do substitution and dynamic quotation have synergistic effects?
- Which one is more beneficial, substitution or dynamic quotation under different parameter settings?
- If we have the flexibility of quoting one of the products dynamically, which one to choose?
- When does substitution or dynamic quotation yield the highest benefits?

No-Substitution		Product 2	
		Static	Dynamic
Product 1	Static	NS-Sd1Sd2	NS-Sd1Dd2
	Dynamic	NS-Dd1Sd2	NS-Dd1Dd2

Substitution		Product 2	
		Static	Dynamic
Product 1	Static	S-Sd1Sd2	S-Sd1Dd2
	Dynamic	S-Dd1Sd2	S-Dd1Dd2

Figure 7: Comparison Scenarios

We use a policy iteration algorithm (coded in C++) to solve (5) as in Bertsekas [4]. A preliminary experiment is conducted to determine the appropriate discretization scheme for lead-time decisions, and we observe that quoting lead-times with a higher precision, i.e., quoting by 0.05 increments compared to 0.025 increments, increases the profit of dynamic quotation with substitution less that 0.1% on average, however nearly quadruples the computational time. Therefore, lead-time quotes are quoted in 0.05 increments in the computational experiments. The average computation time using the policy iteration method is 18.04 minutes. In Section 2.4, we derive valid cuts on the feasible region depending on an endogenous variable, $v^*(x_1, x_2)$. The average computation time decreased to 7.63 seconds over 1612 instances, and the coefficient of variation of the average computation time also decreased from 2.19 to 1.38 by using Proposition 2.

The buffer size for the numerical analysis is selected sufficiently high to minimize the probability of customer rejections due to buffer size, i.e., none of the incoming customer orders are accepted when there are N orders in the system. Buffer size, N, is calculated by assuming that the manufacturer quotes the minimum lead-times for both products and customers place orders for their preferred products. Then, for each product the buffer size is calculated according to $M/M/1/N_i$ queue limiting probability for state N_i such that the limiting probability $(\frac{\lambda}{\mu})^{N_i}(1-\frac{\lambda}{\mu})$ is less than 5x10^{-4} for each product and the maximum of N_i is used as N for both products Ross [45]. The optimality equation for finite state space is given in Appendix A.2.

Some of the parameters are selected as in the numerical analysis of Savasaneril et al. [48], who test the benefits of high precision dynamic lead-time quotation. In all experiments, the lateness cost per item per unit time, $l_1 = l_2 = 1.5$ and service rates, $\mu_1 = \mu_2 = 0.5$ are held constant. Other parameters are selected in order to measure the impact of parameters on profit improvements.

2.5.1 Comparison Scenarios

In this section, we explore the scenarios summarized in Figure 7 in some detail.

1. No-substitution: We use this case as a benchmark to assess the benefit of substitution between two products (or capacities). In this case, we assume that substitution is not possible and we offer product 1 to only customer type 1 and product 2 to only customer type 2. Two independent problems are solved for each product and states are hence one-dimensional.

   The optimality equations for the dynamic quotation of product $i = 1, 2$ is as follows for all states, x_i:

$$\frac{g_i^*}{\nu_i} + v^*(x_i) = \max_{d_i \in (0, d_i^{\max}]} \left\{ \frac{\lambda_i}{\nu_i} p_i(d_i)(r_i - l_i L(x_i, d_i) + v^*(x_i + 1)) + \frac{\mu_i}{\nu_i} v^*((x_i - 1)^+) + \frac{\lambda_i}{\nu_i}(1 - p_i(d_i))v^*(x_i) \right\}, \tag{17}$$

where $\nu_i = \lambda_i + \mu_i$.

When defining the acceptance probabilities under the no-substitution assumption, we ignore the preference of customer of product 1 for product 2. If quoted a zerolead-time, then the customer will accept with probability 1 and if quoted $d^{\max}$ then customer will reject with probability 1. This is also equivalent to quoting the maximum lead-time for other product, i.e., $p_{ii}(d_i, d_j^{\max}) = p_i(d_i)$.

$$p_i(d_i) = -\frac{d_i}{d_i^{\max}} + 1. \tag{18}$$

2. Substitution:

   (a) Static-dynamic: The lead-time quotation is dynamic only for one of the products. For the product quoted static, the lead-time that maximizes the profit is chosen among all possible values solving the model where products are substitutable. The formulations where d_2 is static is given below, similarly the version where d_1 is static can be obtained.
   In this version, d_2 is static, i.e., d_2^S. The below problem is solved for all possible values of $d_2^S \in (0, d_2^{\max}]$ with an appropriate discretization scheme, and the one with the maximum profit is chosen and defined as d_2^{S*}.

$$\begin{aligned} \frac{g_{S2}^*}{\nu} + v^*(x_1, x_2) = &\max_{d_1 \in (0, d_1^{\max}]} \left\{ \left(\frac{\lambda_1}{\nu} p_{10}(d_1, d_2^S) + \frac{\lambda_2}{\nu} p_{20}(d_1, d_2^S) \right) v^*(x_1, x_2) \right. \\ + &\left(\frac{\lambda_1}{\nu} p_{11}(d_1, d_2^S) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2^S) \right) (r_1 - l_1 L(x_1, d_1) + v^*(x_1 + 1, x_2)) \\ + &\left(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2^S) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2^S) \right) (r_2 - l_2 L(x_2, d_2^S) + v^*(x_1, x_2 + 1)) \\ + &\left. \frac{\mu_1}{\nu} v^*((x_1 - 1)^+, x_2) + \frac{\mu_2}{\nu} v^*(x_1, (x_2 - 1)^+) \right\}, \end{aligned} \tag{19}$$

where $\nu = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$ and choice probabilities are defined as in the original version of the problem.

(b) Both static: The lead-time quotation is static for both products (i.e., state-independent). The problem in Equation (20) is solved for every (d_1, d_2) pair for $d_1 \in (0, d_1^{\max}]$ and $d_2 \in (0, d_2^{\max}]$, then the one with maximum profit determines the optimal lead-time quotes for both-static version, i.e., (d_1^{SS}, d_2^{SS}).

$$\begin{aligned} \frac{g^*_{SS}}{\nu} + v^*(x_1, x_2) = &\Bigg\{ \left(\frac{\lambda_1}{\nu} p_{10}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{20}(d_1, d_2) \right) v^*(x_1, x_2) \\ + &\left(\frac{\lambda_1}{\nu} p_{11}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2) \right) (r_1 - l_1 L(x_1, d_1) + v^*(x_1 + 1, x_2)) \\ + &\left(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2) \right) (r_2 - l_2 L(x_2, d_2) + v^*(x_1, x_2 + 1)) \\ + &\frac{\mu_1}{\nu} v^*((x_1 - 1)^+, x_2) + \frac{\mu_2}{\nu} v^*(x_1, (x_2 - 1)^+) \Bigg\}, \end{aligned} \tag{20}$$

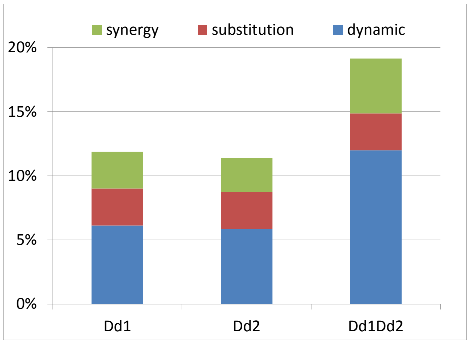
where $\nu = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$ and choice probabilities are defined as in the original version of the problem.

2.5.2 Synergy of Substitution and Dynamic Quotation

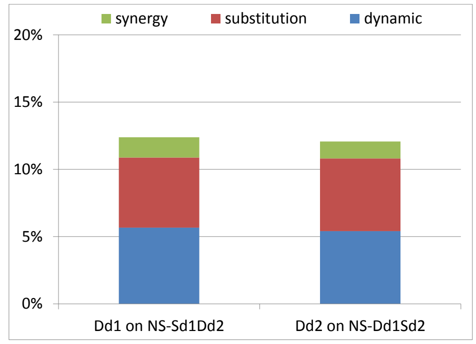
In this section, we explore the synergistic effects of substitution and dynamic quotation. We compare the benefits obtained by dynamic quotation and substitution individually on a base scenario, with dynamic quotation and substitution applied together on a base scenario.

Observation 3 *Higher benefits can be obtained by applying dynamic quotation and substitution together.*

Observation 3 is tested in five cases as shown in Figure 8. The first three cases are improvements obtained with respect to NS-Sd1Sd2 by dynamic quotation of: (i)d_1, (ii)d_2, (iii) both d_1 and d_2; and substitution compared to dynamic quotation and substitution applied together. The last two cases are improvements obtained on



(a) on NS-Sd1Sd2



(b) on NS-Sd1Dd2 and NS-Dd1Sd2

Figure 8: Synergy effects for substitution and dynamic quotation

NS-Sd1Dd2 and NS-Dd1Sd2, respectively. The benefit of substitution and dynamic quotation together is higher than the sum of applying each of them separately in all cases on average, considering the 1612 instances. Note that the benefits obtained by substitution and dynamic quotation of both products on NS-Sd1Sd2 is 19.14% averaged over all instances.

Figure 9 depicts the histogram of percent improvement obtained by synergy of dynamic quotation and substitution for five cases defined above. Note that the acceptance probability for no-substitution case assumes the highest lead-time is quoted for the other product, which may underestimate or overestimate the effect of substitution. We observe that substitution and dynamic quotation does not have synergistic effects in higher number of instances for NS-Sd1Dd2 and NS-Dd1Sd2. Note that in these cases, the profit improvement obtained by substitution is higher compared to other cases. Also, the synergy effects are higher when both products are quoted dynamic, since substitution can be used more effectively with dynamic quotation compared to static quotation.

Observation 4 *Dynamic quotation of both products yields higher benefits than the dynamic quotation of one of the products in all instances.*

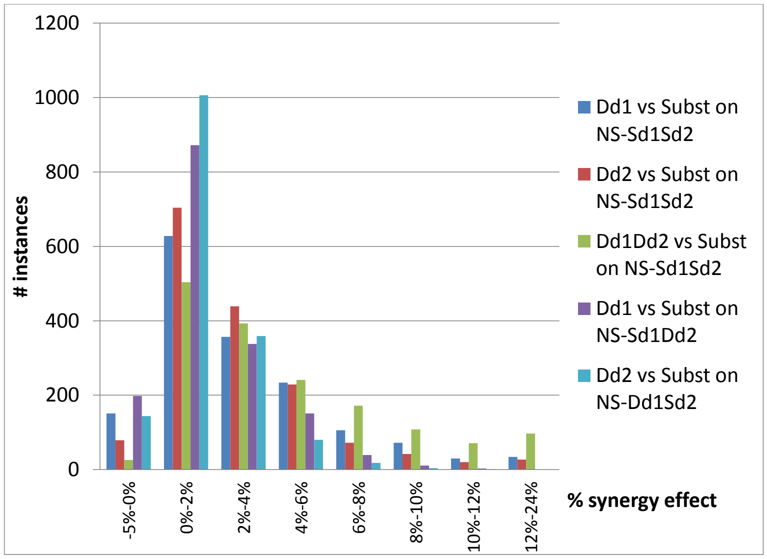


Figure 9: Histogram of percent improvement obtained by synergy of dynamic quotation and substitution

Observation 5 *The benefit of substitution is higher when one of the products is quoted dynamically, compared to the static quotation of both products in all instances.*

Observation 4 can be seen from Figure 8 part (a). Dynamic quotation of both products yields 11.98% compared to 6.13% and 5.86% in case of dynamic quotation of d_1 and d_2 respectively. The benefits obtained by substitution on NS-Sd1Sd2, 2.88% increases to 5.21% on NS-Sd1Dd2 and 5.4% on NS-Dd1Sd2. This is an intuitive result since substitution can be used more efficiently with dynamic quotation.

2.5.3 Impact of Parameters

In this section, we analyze the impact of parameters on improvements over NS-Sd1Sd2. We used different parameter settings to investigate the effect of parameters. Let us first summarize the numerical analysis settings.

The numerical analysis settings are summarized in Table 1 for impact of revenues. We test three levels of $\lambda_1 + \lambda_2$ to represent low, medium, high traffic intensity, two levels of $d^{\max}$ to represent low and high lead-time sensitivity for both products and three levels for each one of the choice probability parameters (where $k_{11} = k_{22}$ and $k_{12} = k_{21}$). Note that customers with $d^{\max} = 4$ are *more* sensitive to lead time quotes than customers with $d^{\max} = 8$ due to the linear choice probability model. The results for $r_2 \in \{4, 5, 10, 15, 20, 25, 30, 35, 40\}$ where revenue of product 1 is held constant at $r_1 = 15$ are reported in Figure 12, where % improvement with respect to NS-Sd1Sd2 are averaged over all other parameters (i.e., $\lambda_1+\lambda_2$, $d_1^{\max} = d_2^{\max}$, $k_{11} = k_{22}$, $k_{12} = k_{21}$).

Table 1: Numerical analysis settings for impact of parameters.

Revenues	Arrival Rates
$r_1 \in \{15\}$ $r_2 \in \{5, 10, 15, 20, 25, 30, 35, 40\}$	$(\lambda_1 + \lambda_2) \in \{0.3, 0.6, 0.9\}$ $\lambda_1/(\lambda_1 + \lambda_2) \in \{50\%\}$
Lead-time Sensitivity	**Choice Parameters**
$d_1^{max} = d_2^{max} \in \{4, 8\}$	$k_{11} = k_{22} \in \{0.15, 0.30, 0.45\}$ $k_{12} = k_{21} \in \{0.15, 0.30, 0.45\}$

The numerical analysis settings for arrival rates and traffic intensity are constructed similarly and given in Tables 2-3.

Table 2: Numerical analysis settings for impact of arrival rates.

Revenues	Arrival Rates
$r_1 = r_2 \in \{5, 15, 25\}$	$(\lambda_1 + \lambda_2) = 0.6$ $\lambda_1 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$
Lead-time sensitivity	**Choice prob. Parameters**
$d_1^{max} = d_2^{max} \in \{4, 8\}$	$k_{11} = k_{22} \in \{0.15, 0.30, 0.45\}$ $k_{12} = k_{21} \in \{0.15, 0.30, 0.45\}$

Table 3: Numerical analysis settings for impact of traffic intensity.

Revenues	Arrival Rates
$r_1 = r_2 \in \{5, 15, 25\}$	$(\lambda_1 + \lambda_2) \in \{0.2, 04, 0.6, 0.8, 0.9\}$ $\lambda_1/(\lambda_1 + \lambda_2) \in \{50\%\}$
Lead-time sensitivity	**Choice prob. Parameters**
$d_1^{max} = d_2^{max} \in \{4, 8\}$	$k_{11} = k_{22} \in \{0.15, 0.30, 0.45\}$ $k_{12} = k_{21} \in \{0.15, 0.30, 0.45\}$

Observation 6 *Dynamic quotation yields higher benefits than substitution when revenues of products are similar. Moreover, substitution yields higher benefits than dynamic quotation as the customer mix gets unbalanced, i.e., the arrival rates of customers differ.*

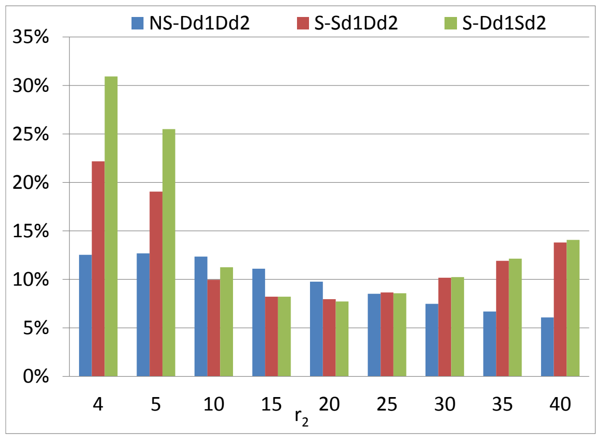
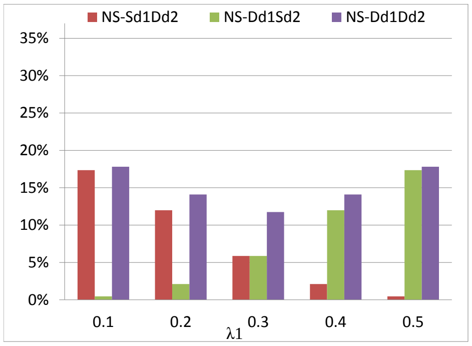
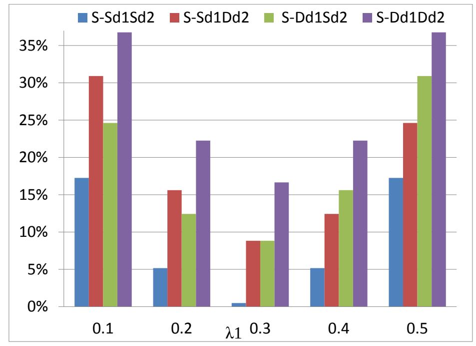


Figure 10: The impact of r_2 on benefits obtained by Dd_1Dd_2 vs. substitution and Dd_1 (Dd_2) on NS-Sd1Sd2

Figure 10 depicts the average improvement values for NS-Dd1Dd2 versus S-Sd1Dd2 and S-Dd1Sd2 in order to compare the benefits obtained by substitution and dynamic quotation when revenues of products differ. We observe that the average improvement in S-Sd1Dd2 and S-Dd1Sd2 are less than that in NS-Dd1Dd2 for the cases $r_2 \in \{10, 15, 20\}$ for parameters given in Table 1. Analyzing these results in depth for each instance, we observe that this observation is valid for $70\% - 80\%$ of the instances for S-Dd1Sd2 and S-Sd1Dd2 scenarios, respectively. Furthermore, if we only take into account the instances with $\lambda_1 = \lambda_2 \in \{0.30, 0.45\}$; in the $92\% - 95\%$ of the instances the observation becomes valid for S-Dd1Sd2 and S-Sd1Dd2 scenarios, respectively. Hence, we can conclude that the observation that the benefit of dynamic quotation is higher than the benefit of substitution when revenues are close, is valid when traffic intensity is not very low.



(a) No-substitution



(b) Substitution

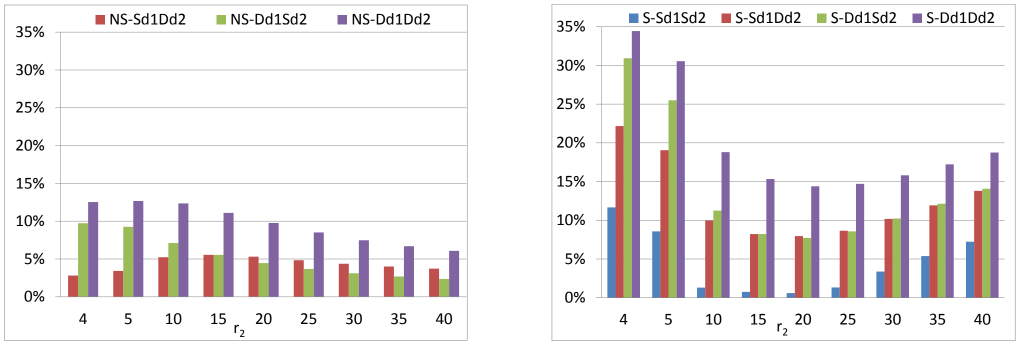
Figure 11: The impact of λ_1 on benefits obtained by Dd_1 and/or Dd_2 w/ or w/o substitution on NS-Sd1Sd2 where $\lambda_1 + \lambda_2 = 0.6$

Observation 6 indicates that if customers have different arrival rates, substitution can be used more effectively, i.e., profit improvements increase much more compared to dynamic quotation as in Figure 11. Figure also depicts that the benefits obtained by substitution and/or quoting both products dynamically increase as the customer mix becomes different in terms of customer types.

Observation 7 *If we have the flexibility of quoting one of the products dynamically, we can obtain higher benefits by quoting dynamically the one with:*

- *higher revenue if there is a revenue difference between products, or*
- *higher arrival rate if arrival rates are different.*

For $\lambda_1 > \lambda_2$, then the profit improvements obtained by NS-Dd1Sd2, and S-Dd1Sd2 are higher than NS-Sd1Dd2 and S-Sd1Dd2, respectively as in Figure 11. If there is an opportunity to quote one of the products dynamically, the one with a higher market share should be preferred. Similarly, Figure 12 depicts that for $r_2 \in \{4, 5, 10\}$, i.e., $r_2 < r_1 = 15$, profit improvements for NS-Dd1Sd2, and S-Dd1Sd2 are higher than NS-Sd1Dd2 and S-Sd1Dd2, respectively.



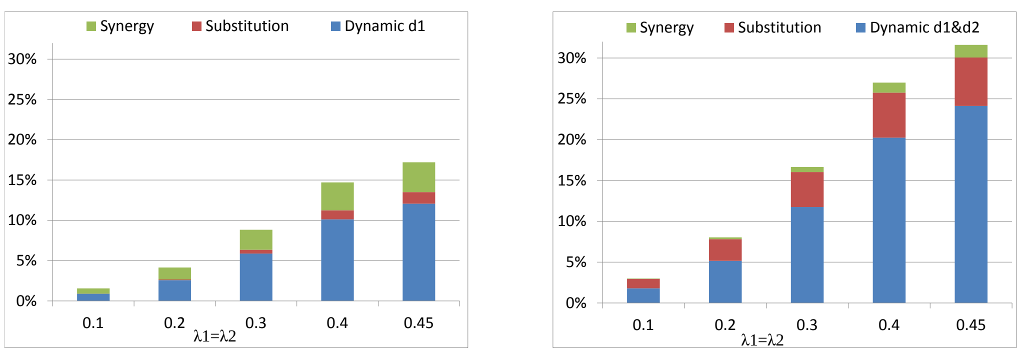
(a) No-substitution

(b) Substitution

Figure 12: The impact of r_2 on benefits obtained by Dd_1 and/or Dd_2 w/ or w/o substitution on NS-Sd1Sd2

Observation 8 *The benefits obtained by substitution and dynamic quotation increases when*

- *revenues differ and the lowest revenue is close to the lateness cost, or*
- *traffic intensity increases, or*
- *arrival rates differ.*



(a) Dynamic d_1 or d_2

(b) Dynamic d_1 and d_2

Figure 13: The impact of $\lambda_1 + \lambda_2$ on benefits obtained by Dd_1 (Dd_1Dd_2), substitution and synergy of both on NS-Sd1Sd2

Figure 12 depicts the improvements obtained by dynamic quotation or substitution and both when revenue of product 2 changes. Substitution and dynamic quotation yields higher benefits when revenues of products differ. As revenues get closer, the profit improvement obtained by offering the product with higher revenue diminishes.

Figure 13 depicts the magnitudes of benefits obtained by substitution, dynamic quotation and the synergistic effects applied on base scenario, NS-Sd1Sd2. The benefits obtained by dynamic quotation of d_1 or d_2 and both d_1 and d_2 increases as total traffic intensity increases. In case of dynamic quotation of only one of the products, the benefits obtained by substitution are lower and synergistic effects are higher compared to the case where both products are quoted lead-times dynamically. When both products are quoted dynamically, more improvement is obtained by substitution as a result of state-dependent lead-times.

2.5.4 Interaction Effects

In this section, we will analyze the interaction effects of parameters that have been analyzed individually in the previous subsection. The interaction of revenue differences ($|r_2 - r_1|$) versus the differences in arrival rates ($|\lambda_2 - \lambda_1|$), total traffic intensity ($\lambda_1 + \lambda_2$) and the sum of cross lead-time elasticities ($k_{11} + k_{12}$) are analyzed as shown in Table 4. The values shown in Table 4 are the percent improvements in profits for S-Dd1Dd2 over the base scenario of NS-Sd1Sd2, averaged over all instances.

Table 4: Revenue differences vs. arrival rate differences, total traffic intensity and the sum of cross lead-time elasticities

		$\|\lambda_2 - \lambda_1\|$			$\lambda_1 + \lambda_2$			$k_{11} + k_{12}$	
		low	high		low	high		low	high
$\|r_2 - r_1\|$	low	16.48%	36.77%		4.46%	29.79%		15.81%	19.26%
	high	21.91%	53.99%		17.63%	29.35%		15.18%	29.02%

1. **Revenue versus Arrival Rate:** The interaction of revenue differences ($r_2 - r_1$) and the differences in arrival rates ($\lambda_2 - \lambda_1$) are analyzed where "low" represents equal revenues/arrival rates for both products, and "high" for revenue differences is $\{-11, -10, 10, 15, 20, 25\}$ and $\{-0.4, 0.4\}$ for the differences in

arrival rates. As differences in revenues or arrival rates of products 1 and 2 increase, the benefits obtained by substitution and dynamic quotation of lead-times increase. Note that the percent improvements in profits rise up to 53.99% when revenue differences and arrival rate differences are high.

2. **Revenue vs. Total Traffic Intensity:** Percent improvements in profits of S-Dd1Dd2 over the base scenario of NS-Sd1Sd2 are analyzed with respect to revenue differences and traffic intensity in Table 4, where "low" traffic intensity values are $\{0.2, 0.3\}$, and "high" traffic intensity values are $\{0.8, 0.9\}$. When traffic intensity is low, increase in differences of revenues increase the percent improvement significantly, i.e., from 4.46% to 17.63%. On the other hand, when traffic intensity is high; as revenue differences change, the change in percent improvements is not significant.

3. **Revenue vs. Cross Lead-time Elasticities:** The interaction of the differences in revenues of products and the sum of cross lead-time elasticities for customer 1 are analyzed together as shown in Table 4. The sum of the cross lead-time elasticities for customer 1, i.e., $k_{11} + k_{12}$, are assumed to be "low" for $\{0.3, 0.4, 0.45\}$, and "high" for $\{0.75, 0.85, 0.95\}$. Table 4 shows that when the sum of cross lead-time elasticities are low, i.e., the choice probability of a product is less sensitive to other products' lead-time, increase in revenue difference does not significantly change the profit improvements. However, when choice probabilities dependence upon other products' lead-time increase, high revenue differences results in almost 10% (19.26% to 29.02%) increase in profit improvements.

2.6 Conclusion

We considered a dynamic lead-time quotation problem in a make-to-order system with two substitutable products with the objective of maximizing the long-run average expected profits. There are two customer types with different choice probabilities for the products. Upon obtaining a lead time quote for each product, the customer chooses to order one of the products or leaves system without placing an order. We define the choice probabilities for each product to be linear in the quoted lead-time and cross lead-time effects. We first analyze the optimal profit and define the regions where it is concave or linear. Then, we analyze the structure of optimal lead-time policy and restrict the optimal lead-time quotes to certain values under certain conditions. We find that if the expected profit (revenue minus late delivery penalties) evaluated at maximum lead-time quote is negative, then customer should be rejected by quoting the maximum lead-time. We observe that the behavior of optimal lead-time quote for product 1 (2) is not monotonic in state of product 2 (1) when lead-times are quoted dynamic and substitution is enabled.

We evaluate the value of substitution and dynamic quotation through a numerical study. We first compare the profit improvements obtained by substitution and/or dynamic quotation applied on base scenario with no-substitution and static quotation. Substitution and dynamic quotation has synergistic effects, i.e., the benefit of applying substitution and dynamic quotation is higher than the sum of the profit improvements obtained only by dynamic quotation and only by substitution. We investigate the impact of several problem parameters (such as revenues, arrival rates, total traffic intensity) and customer preferences (measured by lead-time sensitivity and choice parameters) on the profit improvements obtained by dynamic quotation and/or substitution. We observe that higher differences in revenues, or arrival rates, higher traffic intensity, lead-time sensitivity or cross lead-time elasticity results in higher profit improvements by dynamic quotation and substitution.

Our model assumes a two-product, two-customer type and Poisson arrivals and departures. These assumptions could be relaxed in several ways. The model can be extended to an n-product, m-customer type problem and analyzing general arrivals and departures. Further the manufacturer may have different processing options, such as regular versus expediting at a higher cost.

CHAPTER III

DYNAMIC LEAD-TIME QUOTATION WITH ORDER CANCELLATIONS

3.1 Introduction

In this chapter, we study a dynamic lead-time quotation problem in a make-to-order (MTO) system with order cancellations. Cohen et al. [12] state that 43 of 143 orders they collected in their empirical study were cancelled in the semiconductor equipment supply manufacturer. Ignoring cancellations may lead to erroneous (non-optimal) lead-time quotes, under-utilization of resources and, in turn, lower profits. We assume that all customers and products are identical and customers may cancel their orders while their orders are waiting in the queue or being processed, or at the time of order delivery. We assume that cancellations occur at a known exponential rate, and hence is dependent on the amount of time that a customer waits for order delivery. On the other hand, cancellations at delivery may depend on other additional reasons such as quality. We formulate the problem as a Markov decision process by characterizing time-in-system, expected tardiness and cancellation probability. Then, we analyze the effects of cancellation parameters on the optimal profit. Through an extensive numerical study, we identify counter-intuitive settings where profits may increase as the cancellation rate increases. Finally, we compare cancellation scenarios with corresponding no-cancellation scenarios in terms of optimal profit.

This chapter is organized as follows. In Section 3.2, we review the related literature. In Section 3.3, we model and analyze the dynamic lead-time quotation problem with order cancellations. In Section 3.4, we discuss the affects of cancellation parameters on the optimal profit and compare cancellation versus no-cancellation scenarios.

We present our major findings and conclude in Section 3.5.

3.2 Literature Review

There are two streams of literature relevant to our research, namely dynamic lead-time quotation and order cancellations. Dynamic lead-time quotation is already reviewed in Chapter 2, hence we review the order cancellation literature in this section.

Cheung and Zhang [10], Yeo and Yuan [55] and Yuan and Cheung [59] consider order cancellations in periodic review inventory models. Cheung and Zhang [10] model and analyze customer order cancellations in a periodic review (s, S) inventory model with Poisson arrivals, deterministic demand lead-times and supply lead-times. The system behavior is analyzed for cases with no setup cost and fixed setup cost. Cancellation fees and the impacts of various conditions of cancellation can be computed using their model. They show that cancellations increase total system costs, and the probability of cancellation and the expected cancellation time affects the magnitude of the cost. They assume a Bernoulli type cancellation behavior in which order cancellations for each demand reservation occur with probability p, and the timing of the cancellation given that a demand will be cancelled is random with density function $f(t)$. Yeo and Yuan [55] consider a single item, periodic review inventory model under supply uncertainty and demand cancellation. They quantify the importance of reducing the variance of either the distribution of yield or the distribution of demand cancellation. They show that a stochastically larger demand cancellation does not imply a reduction in the optimal cost. Furthermore, they develop a bound on the difference between the optimal cost in the presence of different cancellation behavior, and the bound is proportional to the difference between the mean number of items not eventually cancelled. Yuan and Cheung [59] consider a periodic inventory model where reservations and cancellations affect the ordering policy. They assume that all demands are reserved by a lead-time of one period and orders can be canceled during the reservation period. They show that the order-up-to policy is optimal whose re-order point is dependent on the reservation parameter. They assume there is no penalty cost for cancellations and in each period the expected amount of canceled orders is a constant.

You [56], You and Wu [58] and You and Hsieh [57] consider joint ordering and pricing problem with cancellations. You [56] considers a joint ordering and pricing problem for a single period model for perishable products sold over a short selling period. Price-dependent demand and customer cancellations are taken into account as an extension to newsboy problem. Ordering, dynamic pricing and over-reservation decisions are addressed simultaneously, in order to maximize the total expected profit. Optimal ordering policy has an order-up-to structure. Moreover, the optimal prices are nondecreasing with the amount of reserved units and nonincreasing with the number of remaining decision periods. You and Wu [58] consider a joint ordering and pricing decision problem with deterministic price-dependent demand and cancellations, with an objective of maximizing total profit over a finite time planning horizon by determining the optimal advance sales price, spot sales price, order size, and replenishment frequency. They assume that a fraction of orders are canceled at a constant rate. They derive closed-form formulas to find the optimal decisions for linear and exponential demand cases. You and Hsieh [57] consider a joint production and pricing problem under the condition that back-order is allowed. They assume that back-ordering customers may cancel their orders and demand cancellations occur at a constant rate. Holding and penalty costs are calculated by a system of differential equations for inventory level, however the impact of cancellation on the optimal cost of the system is not considered.

Cohen et al. [12] and So and Zheng [51] study impact of cancellations for semiconductor manufacturers. Cohen et al. [12] measure how the semiconductor equipment supplier in their study perceived the cost of cancellation and holding relative to the cost of late shipment. They conclude that the supplier is very conservative when commencing the order fulfillment based on soft orders in fear of holding costs and order cancellations which are perceived as much more important relative to the cost of the delay. So and Zheng [51] analyze how the supplier's variable lead-times and the retailer's forecast demand updating can contribute to the high degree of order quantity variability experienced by semiconductor manufacturer. Order cancellations, which are modeled through negatively correlated external demands, also increase the order fluctuations. They conclude that maintaining consistent delivery lead-times especially for products with highly variable and auto-correlated demands, or during the low or high demand season becomes important to reduce the order fluctuations.

Gayon et al. [22] consider a make-to-stock (MTS) supplier producing a single item, and serving multiple customer classes with limited capacity. Some customer classes provide advanced demand information (ADI) which is not perfect, and customers may change the expected due date, or cancel the order. The optimal production policy is a base-stock policy with state-dependent base-stock levels, and the optimal inventory allocation policy is a state-dependent rationing policy such that fulfilling orders from a particular class is optimal only if the inventory level is above the rationing level of that class.

Rubino and Ata [46] consider a dynamic control problem for a MTO system with parallel servers, but they consider outsourcing and resource allocation decisions instead of lead-time quotes by making heavy traffic approximations to minimize long-run average costs by dynamically making outsourcing and resource allocation decisions. Customers have rigid due-dates, therefore the system manager may outsource orders to meet these constraints and also customers may cancel their order subject to a cancellation penalty. They approximate the problem by a Brownian control problem making heavy traffic approximations and propose a nongreedy policy for the general parallel server system.

In our work, we investigate the effect of cancellations on dynamic lead-time quotation, and present results on effects of cancellation parameters on the optimal profit and comparison of cancellation vs. no-cancellation scenarios. We show that increases in cancellation rate do not always imply a decrease in the optimal profit and through a numerical study we investigate the cases in which the optimal profit increases as cancellation rate increases.

3.3 Model Formulation

We model a MTO system where all customers and items are identical, and customers place orders one at a time. Customers arriving according to a Poisson process with rate λ are served by a single server with service times exponentially distributed with rate μ. An arriving customer is quoted a lead-time, d, and decides to place an order with probability $f(d)$, where $f(d)$ is a decreasing function of $d \in [0, d_{max}]$. A unit lateness cost of l is incurred per unit time when an order is delivered later than the quoted lead-time. Expected tardiness duration of an order is denoted as $L(i, d)$ if d is the quoted lead-time and i is the number of orders in the system, which denotes the system state. In state i, the expected profit of $r - lL(i, d)$ is obtained from a customer if customer accepts the quoted lead-time d and places order and also does not cancel the order while waiting in the queue, during processing or at delivery.

$L(i, d)$ is calculated assuming the system operates under the first-come first-served (FCFS) discipline as follows:

$$L(i, d) = \int_{d}^{\infty} (y - d) g_i(y) dy, \tag{21}$$

where $g_i(.)$ denotes the probability density funciton (pdf) of time in system of an incoming order when there are i orders in the system.

There are two main factors that define the nature of the system with cancellations: (i) the probability that an order will be cancelled, and (ii)the timing of the cancellation. Cancellations may occur at any time while in system, i.e., while in queue or being processed; hence, we assume that each order may be cancelled after waiting an exponentially distributed amount of time and each cancellation takes place independently of all other cancellations, service completions, and arriving jobs. If an order is cancelled while in process, we assume that the manufacturer stops processing and continues with the next order in the queue. On the other hand, if an order is cancelled before its processing starts, then it is simply removed from the queue. We assume that the customer is charged an aggregate average penalty of p for a cancellation. p may be either positive (resulting in a profit for the manufacturer) or negative (resulting in a cost for the manufacturer). A cancellation (return) can also take place after an order's processing is completed, namely, at the time of delivery. Since the order processing is already completed, the manufacturer can salvage the product in this case, and this salvage value, s, which includes revenue due to cancellation penalty and salvage costs, may be either positive or negative.

We assume that each order in the queue or in proces is cancelled at an exponential rate of γ. Hence, the probability of cancellation before delivery depends on the cancellation rate, service rate and the number of orders in the system. We drop μ and γ from notation, and simply use $q(i)$ to denote the probability of order cancellation prior to delivery. We assume that cancellations at delivery occur with probability q.

We formulate the dynamic lead-time quotation problem as an infinite-horizon Markov decision process (MDP). State $i(t)$ denotes the number of orders in the system at time t. We drop t from notation, and simply use i to denote the state. The manufacturer controls the demand rate and therefore the profit of the system, by quoting lead-times, $d(i)$, when the system is in state i. Note that the effective cancellation rate at state i is equal to $i\gamma$ which results in nonuniform transition rates. To be able to use uniformization, we assume a finite buffer size N, which essentially means that none of the incoming customer orders are accepted when there are N orders in the system. We assume a sufficiently high value for N in the numerical analysis, so that the probability of customer rejection due to a full buffer is negligibly small.

Let $h(i)$ denote the relative value function of being in state i and g denote the expected average profit per unit time. The continuous-time model is converted to an equivalent discrete-time model through uniformization (Lippman [36]) and 22 denotes the corresponding optimality equation for $i = 0, 1, ..., N-1$.

$$\begin{aligned}\frac{g^*}{\nu} + h^*(i) \quad &= \quad \max_{d \in [0, d^{\max}]} \Bigg\{ \frac{\lambda}{\nu} f(d) \Bigg(q(i)p + (1-q(i))qs \\ &+ \quad (1-q(i))(1-q)(r - lL(i,d)) + h^*(i+1) \Bigg) + \frac{\lambda}{\nu}(1-f(d))h^*(i) \\ &+ \quad \frac{\mu}{\nu} h^*((i-1)^+) + \frac{i\gamma}{\nu} h^*((i-1)^+) + \frac{(N-i)\gamma}{\nu} h^*(i) \Bigg\}, \qquad (22)\end{aligned}$$

where $\nu = \lambda + \mu + N\gamma$ is the uniformization rate, and $(.)^+$ denotes $\mathrm{Max}(\cdot, 0)$.

At state i, the customer accepts the quote with probability $f(d)$ and places an order, increasing the system state to $i + 1$. Expected profit depends on if and when the order is cancelled. If the order is cancelled while in queue or in proces, which occurs with probability $q(i)$, then the "profit" (which can be negative) is p. If the order is cancelled at delivery with probability q, then the "profit" is s. If the order is not cancelled, the expected profit becomes $r - lL(i, d)$. The customer rejects the quote with probability $1-f(d)$, and the state remains the same. When the processing of an order is completed, the system state decreases by one to $i - 1$. At a decision epoch, the probability that an order is cancelled is $(i\gamma)/\nu$ and changes the state to $i - 1$. The last term refers to the fictitious transitions from the state to itself for uniformization, and no change occurs in system state (Bertsekas [4]).

For $i = N$, the optimality equation can be written as:

$$\frac{g^*}{\nu} + h^*(N) = \frac{\lambda}{\nu} h^*(N) + \frac{\mu + N\gamma}{\nu} h^*(N-1), \qquad (23)$$

where $h^*(N) = 0$.

We derive the closed-form expressions for the cancellation probability, $q(i)$, and expected tardiness, $L(i,d)$, in the following lemmas.

Lemma 4 *The closed-form expression for $q(i)$, the probability that an incoming order will be cancelled before service completion given i orders in the system upon arrival, is:*

$$q(i) = \frac{(i+1)\gamma}{\mu + (i+1)\gamma}. \tag{24}$$

Proof. Proof. The proofs of all propositions, lemmas and theorems are included in Appendix B. ■

After plugging in the expression for $q(i)$ and rearranging the terms, the optimality equation reduces to the following.

$$\frac{g^*}{\nu} + h^*(i) = \max_{d\in([0,d^{\max}]} \Bigg\{ \frac{\lambda + (N-i)\gamma}{\nu} h^*(i) + \frac{\mu + i\gamma}{\nu} h^*((i-1)^+) + \frac{\lambda}{\nu} f(d)$$
$$\left(p + \frac{\mu}{\mu + (i+1)\gamma}(s - p + (1-q)(r - s - lL(i,d)) + h^*(i+1) - h^*(i)\right)\Bigg\}. \tag{25}$$

Lemma 5 *The closed-form expression for $L(i,d)$ is:*

$$L(i,d) = \frac{1}{\gamma^i i!} \prod_{k=0}^{i} (\mu + k\gamma) \sum_{k=0}^{i} (-1)^k \binom{i}{k} \frac{e^{-(\mu+k\gamma)d}}{(\mu + k\gamma)^2}. \tag{26}$$

We show how an optimal policy changes according to problem parameters in the following theorems. We first prove properties of $L(i,d)$ with respect to i and d, which are used in our theoretical analysis. Then, we show that the value functions are non-increasing in i.

Lemma 6 *$L(i,d)$ is strictly decreasing and convex in $d \geq 0$, and strictly increasing in i.*

Lemma 6 shows that the expected tardiness duration under order cancellations has the same properties as in the no-cancellation case. Using the properties of $L(i,d)$ in Lemma 7, we show that the relative value functions are non-increasing in state variable i.

Lemma 7 *The relative value functions under an optimal policy, i.e., $h^*(i)$, are non-increasing in i.*

Next, we show that the optimal lead-time quotes in state i are confined within the region that is defined as:

$$\Psi_i = \{d \in [0, d^{\max}] : \alpha_i(d) \geq 0\}. \tag{27}$$

where $\alpha_i(d) = p + \frac{\mu}{\mu+(i+1)\gamma}(s - p + (1-q)(r - s - lL(i,d))$.

$\alpha_i(d)$ is the expected profit given that the order is placed. $\alpha_i(d)$ is non-increasing in i since $L(i,d)$ is strictly increasing in i by Lemma 6 and $\frac{\mu}{\mu+(i+1)\gamma} \geq \frac{\mu}{\mu+(i+2)\gamma}$. Hence, we can conclude that there exists a threshold state $\bar{i}$ such that $d^*(i) = d^{\max}$ for $i \geq \bar{i}$.

Proposition 3 *If $\Psi_i = \emptyset$, then $d^*(i) = d^{\max}$. Otherwise, $d^*(i) \in \Psi_i$.*

The manufacturer "rejects" customers by quoting $d^{\max}$ if the system is already full and accepting a customer could lead to high lateness penalties and negative profits.

Theorem 2 *All other parameters being equal, as the profits of orders that are not cancelled, cancelled while in queue or in process, or cancelled at delivery increase, i.e., r, p or s, respectively, then the optimal expected average profit per unit time increases.*

In Theorem 2, we show that as the profits of orders - whether cancelled at any time or not- increase, then the optimal expected average profit per unit time increases. However, the impact of each parameter on profits is different. For example, experimental results in Section 3.4 show that one unit increase in p increases g^* more than a unit increase in s. We also compare cancellation and no-cancellation scenarios, and find the corresponding cancellation penalties that make a cancellation scenario at least as good as a no-cancellation scenario in terms of the optimal expected average profit per unit time.

Table 5: General parameter setting.

λ	r	$d^{\max}$
$\{0.15, 0.3, 0.45, 0.6, 0.75, 0.9, 0.99\}$	$\{5, 7.5, 10, 15, 25\}$	$\{4, 8\}$

Theorem 3 *All other parameters being equal, as the cancellation probabillity at delivery, i.e., q, increases; then the optimal expected average profit per unit time decreases.*

In Theorem 3, we show that as cancellation probability at delivery increases, the optimal expected average profit per unit time decreases. However, a similar argument does not hold for the cancellations occurring while in queue or processing, i.e, as γ increases. Using an experimental study, we show that as cancellation rate, γ increases, the optimal expected average profit per unit time also increases for certain settings.

3.4 Computational Study

In this section, using a numerical study we analyze the optimal expected average profit per unit time and the optimal lead-time quotes under different problem parameters. We also identify the cancellation parameters that result in higher optimal expected average profit per unit time compared to no-cancellation.

In the numerical study, we choose the parameters that are not cancellation-related (λ, r, $d^{\max}$, l, and μ) as in Savasaneril et al. [48], who test the benefits of high-precision dynamic lead-time quotation. We consider three additional parameters that affect the system behavior which are not cancellation-related: (i) revenue per customer, r; (ii) customer arrival rate, λ; (iii) sensitivity to the quoted lead-times, which can be measured by $d^{\max}$. If $f_1(d) \geq f_2(d)$ for all $d \in [0, d^{\max}]$, and $f_1(d) > f_2(d)$ for some $d \in [0, d^{\max}]$, then customers under $f_2(d)$ are *more sensitive* than customers under $f_1(d)$. In all experiments, the lateness cost per unit time and service rate are held constant, $l = 1.5$ and $\mu = 1$.

We use a relative value iteration algorithm to solve the optimality equation as in Bertsekas [4]. The action space is discretized with 0.01 increments. We conduct a

Table 6: Cancellation-related parameter setting.

λ/γ	q
{ 2, 3, 4, 5 }	{0, 0.05, 0.1, 0.15, 0.2, 0.25}
r/p	r/s
{-3, -2, -1.5, 1.5, 2, 3}	{-3, -2, -1.5, 1.5, 2, 3}

Table 7: Acceptance probability functions.

Type	$d^{\max}$	Function $f(d)$
linear	4	$1-(d/4)$
	8	$1-(d/8)$
convex	4	$1-(d/4)^{1/4}$
	8	$1-5/8d$, for $0 \leq d \leq 1$
		$3/7(1-d/8)$, for $1 \leq d \leq 8$
concave	4	$1-(d/4)^4$
	8	$1-(d/8)^4$

preliminary experiment to determine the appropriate buffer size, N. We observe that increasing the buffer size from 20 to 30 affects the optimal expected average profit per unit time less than 0.01%, but doubles the computational times. Hence, we set $N = 20$.

We consider six different acceptance probability functions (as in Savasaneril et al. [48]) that capture two dimensions about the customers: (i) sensitivity to the quoted lead-times as given in Table 5, and (ii) early versus late sensitivity by considering different types (i.e., linear, convex, concave) of acceptance probability functions. In convex (concave) $f(d)$, a slight increase in the lead-time quote results in a significant decrease in acceptance probability for small (large) values of d, hence we call convex (concave) functions *early (late) sensitive*. Figure 14 depicts linear, convex and concave acceptance probability functions given in Table 7.

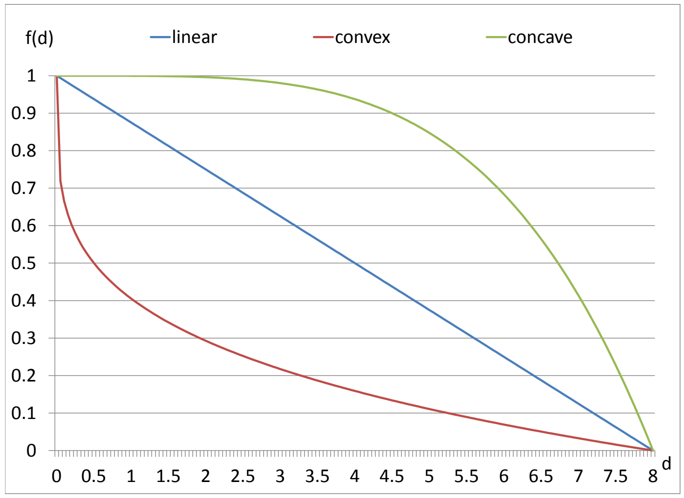
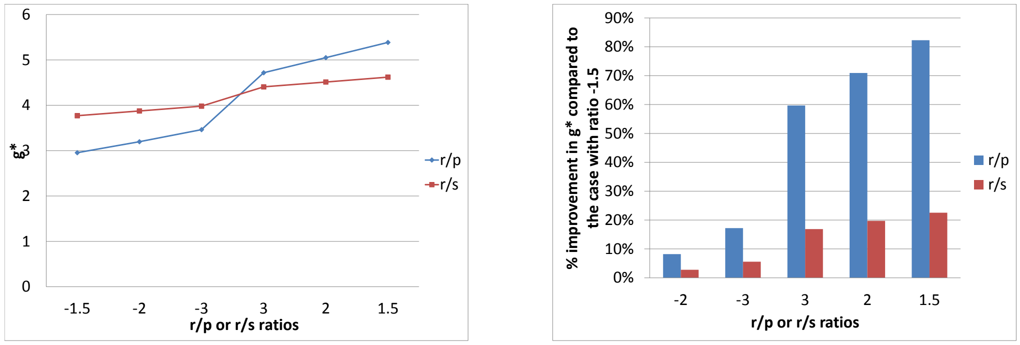


Figure 14: Acceptance probability functions

3.4.1 Analysis of Dynamic Lead-time Quotation Under Order Cancellations

In this section, we analyze the impact of cancellation parameters on the optimal expected average profit per unit time, g^*, and the optimal average lead-time quotes, d^*. Average lead-time quotes are calculated by using the limiting probability distributions and averaging over all values of i using these limiting probabilities. The change in g^* and percent improvement obtained by increasing p or s for $r/p \in \{-1.5, -2, -3, 3, 2, 1.5\}$ and $r/s \in \{-1.5, -2, -3, 3, 2, 1.5\}$ are reported in Figure 15. The reported values are averaged over all other parameters and acceptance probability functions which are summarized in Tables 5 - 7. Note that r/p and r/s are ordered so that p and s increase. Figure 15 confirms the result of Theorem 2, g^* increases as p or s increases. Percent change in g^* depicts the percent change on the lowest p or s values, i.e., r/p or r/s ratio is -1.5.



(a) g^* (b) Percent change in g^*

Figure 15: Change in average g^* as r/p and r/s changes averaged over all other parameters and acceptance probability functions.

Observation 9 *Optimal expected average profit per unit time is more sensitive to the changes in p than s.*

Observation 9 indicates that changes in p impact g^* more than the changes in s. Assuming $r = 15$, increasing s from -10 (corresponds to $r/s = -1.5$) to 10 results in a 22% increase in g^*, whereas the same change in p results in a 82% increase in g^* as observed in Figure 15.

Figure 16 depicts the changes in the average optimal lead-time quotes, d^*, for $r/p \in \{-1.5, -2, -3, 3, 2, 1.5\}$ and $r/s \in \{-1.5, -2, -3, 3, 2, 1.5\}$, where the reported values are averaged over all other parameters and acceptance probability functions which are summarized in Tables 5-7.

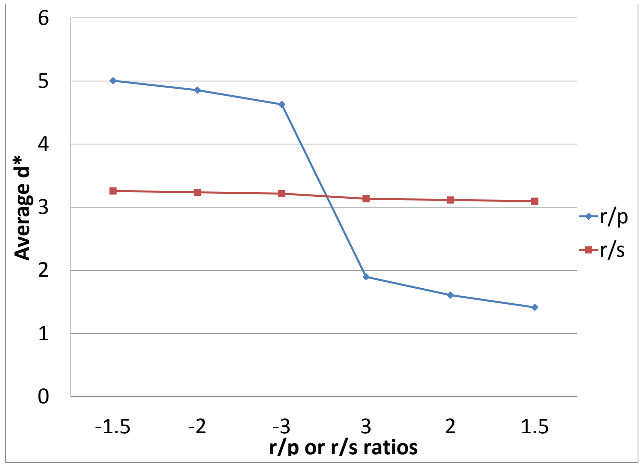


Figure 16: Optimal lead-time quote as r/p and r/s change, averaged over all other parameters and acceptance probability functions.

Observation 10 *Average optimal lead-time quotes decrease as p or s increases. As p shifts from negative to positive, average optimal lead-time quotes significantly decrease.*

As p or s increases, accepting a customer to the system by quoting a lower lead-time becomes more profitable. Hence, the average optimal lead-time quotes decrease as s or p increases. Average optimal lead-time quotes do not significantly change as s increases, since cancellations at delivery do not affect the system state. On the other hand, as p increases, average d^* decreases significantly. A dramatic decrease occurs when p switches from negative to positive values as can be seen in Figure 16. Hence, we can conclude that an increase in p translates into accepting more customers to the system and a higher profit as a result.

In Theorem 3, we show that as the cancellation probability at delivery, q, increases, then the optimal expected average profit per unit time decreases. However, we cannot prove a similar result for cancellations occurring while in queue or in process. An increase in the cancellation rate, γ, does not always result in a decrease in g^*. We test four levels of γ, $\lambda/\gamma \in \{2, 3, 4, 5\}$ for each arrival rate value. In order to identify the settings in which cancellations are beneficial, all other parameters being equal we decrease λ/γ from 5 to 4, 4 to 3, and 3 to 2, and observe that g^* either increases or decreases in these 3 cases. Hence, γ in conjuction with other parameters determine whether cancellations are beneficial or not. Figure 17 shows an example where $\lambda = 0.9$, $r = 5$, $q = 0$, $d^{\max} = 4$ and $r/s = 1.5$, and the optimal expected average profit per unit time increases and decreases, respectively, as γ increases depending on the choice of r/p. For $r/p \in \{1.5, 2\}$, as the cancellation rate increases, profit increases. On the other hand, for the remaining r/p values, as the cancellation rate increases, the profit decreases. We observe that g^* increases in 15%, 22% and 5% of the cases, respectively, for linear, convex and concave acceptance probability functions as γ increases.

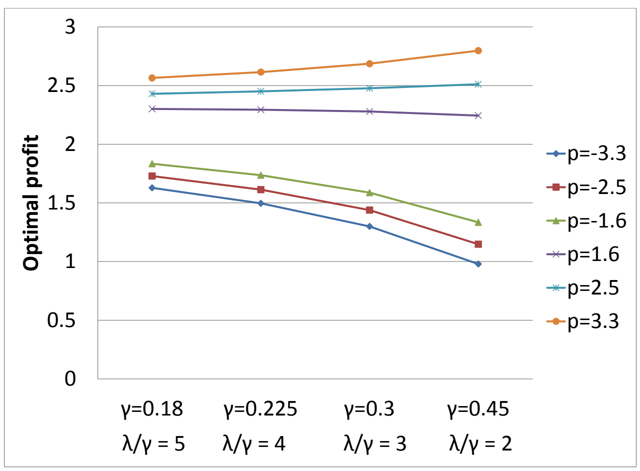
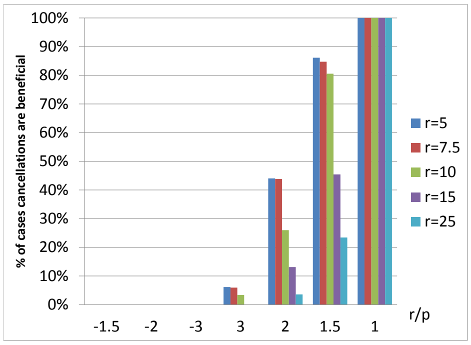
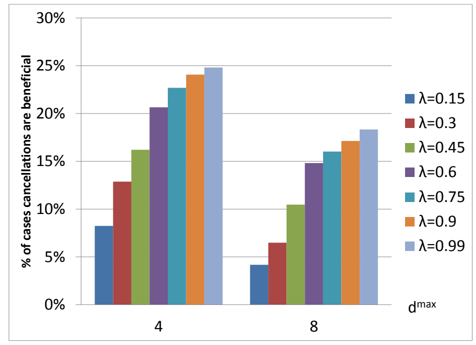


Figure 17: An example of the cases where cancellations are beneficial and not beneficial under different r/p values, as γ increases.

Observation 11 *Optimal expected average profit per unit time increases as the cancellation rate, γ, increases when (i) p is positive and high, (ii) r is low, (iii) λ is high, and (iv) d^{max} is low. In these cases, the average optimal lead-time quotes are lower compared to the cases where the optimal expected average profit per unit time decreases as the cancellation rate increases.*



(a)



(b)

Figure 18: Percentage of cases where g^* increases as γ increases for linear acceptance probability functions.

An investigation of cases where g^* increases as λ/γ decreases from 5 to 4, 4 to 3, and 3 to 2 reveals the results summarized in Observation 11 and depicted in Figure 18. Cancellation profits must be positive in order to observe an increase in the optimal expected average profit per unit time and more such instances are observed for higher penalty values as expected. Revenue from an order being low implies, all profits and costs are closer to each other since we set $l = 1.5$; and it becomes more likely to make cancellations more profitable. Higher traffic intensity and customers who are more sensitive to the quoted lead-times definitely contribute to an increase in the optimal expected average profit per unit time as the cancellation rate increases. When customers are more sensitive to the quoted lead-times and traffic intensity is high, by quoting lower lead-times the manufacturer accepts more customers to the system, and most of these orders are cancelled and generates higher profit compared to rejecting customers. Figure 19 reveals that the average optimal lead-time quotes are lower in cases where the optimal expected average profit per unit time increases as γ increases. Hence, more customers are accepted to the system. Moreover, average of the optimal lead-time quotes are the highest for concave and the lowest for convex acceptance probability functions. Customers are late sensitive in concave cases so increasing lead-time quotes to a higher level - compared to convex and linear cases- affect the acceptance probability less than other types of acceptance probability functions.

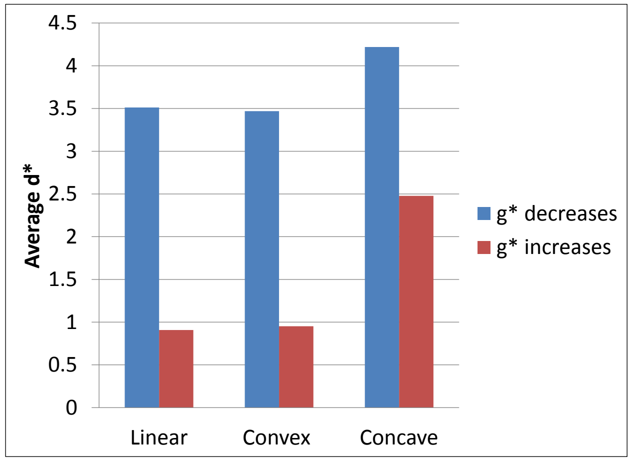


Figure 19: Average optimal lead-time quotes for cases where g^* increases and decreases as γ increases.

3.4.2 Comparison of Cancellation and No-Cancellation Scenarios

In this section, we compare the optimal expected average profit per unit time of cancellation scenarios for cancellation parameters given in Table 6, with a corresponding no-cancellation scenario by setting γ and q to zero. In Section 3.4.1, we observed that an increase in the cancellation rate, γ, does not always lead to a decrease in the optimal expected average profit per unit time. Hence, we identify the cases for which there exists a better scenario with cancellation in terms of profit.

We compare each no-cancellation scenario with parameters as given in Table 5 with corresponding cancellation scenarios with cancellation parameters given in Table 6 for different types of acceptance probability functions and quotation strategies, i.e., dynamic and static quotation. Lead-time quotes are state independent in static quotation. Figure 20 shows that for 50%, 67%, and 83% of no-cancellation scenarios, there exists a cancellation scenario, respectively, for concave, linear and convex acceptance probability functions with higher optimal expected average profit per unit time. We also compare the optimal expected average profit per unit time for static lead-time quotation and observe that higher percentages of correponding cancellation scenarios can be generated for no-cancellation scenarios. In static quotation case, the manufacturer cannot control the acceptance of customers on state basis. Hence, cancellations with several cancellation probabilities and profits enable the manufacturer to achieve higher profits.

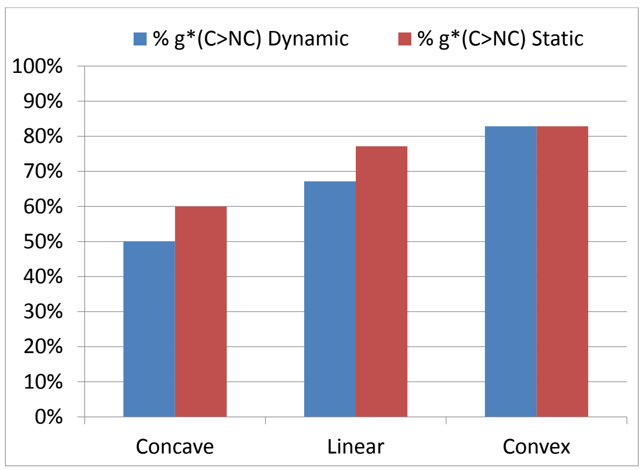


Figure 20: Percentage of no-cancellation scenarios for which a corresponding cancellation scenario can be generated with cancellation parameters given in Table 6.

3.5 Conclusion

We considered a manufacturer's dynamic lead-time quotation problem in a MTO system considering order cancellations in two ways: (i) cancellations occur while an order is in queue or in process, and (ii) cancellations occur after the order is processed. We model the dynamic lead-time quotation problem using a Markov decision process model over an infinite horizon, with the objective of maximizing the long-run average expected profit per unit time. We first derive the expressions for time-in-system, expected tardiness duration and cancellation probability for cancellations occurring while the order is in queue or in process. Then, we analyze the structure of tardiness duration and relative value functions which enable us to define the set of the optimal lead-time quotes and show that the optimal expected average profit per unit time increases as the profits of orders that are not cancelled, cancelled while in queue or in process and cancelled at delivery increase.

Through an extensive numerical study, we observe that the optimal expected average profit per unit time and the average optimal lead-time quotes are more sensitive to the changes in the profit of orders that are cancelled while in queue or in process than changes in the profit of orders that are cancelled at delivery. An increase in the cancellation rate does not always translate into a decrease in the optimal expected average profit per unit time. When the cancellation profit is positive and high, the revenue from an order is low, the arrival rate is high, and the lead-time sensitivity is high; it is more likely that the optimal expected average profit per unit time increases as the cancellation rate increases. Moreover, average optimal lead-time quotes are lower in cases where the optimal expected average profit per unit time increases as the cancellation rate increases. Lastly, we compare the cancellation scenarios and the corresponding no-cancellation scenarios. Results show that under certain parameter settings, there exists a corresponding cancellation scenario which is better than a no-cancellation scenario in terms of the optimal expected average profit per unit time.

Our model assumes a single-item, and a single-customer type. These assumptions can be relaxed by addition of different customer classes with different revenues, and cancellation probabilities to the model. Lead-time quotes at any state may be different for these different types of customer classes to reflect the choice of customer type for manufacturer at that time according to revenue and cancellation probability of customers. Potential future work also includes studying the model with cancellation rates depending on the time-in-system.

CHAPTER IV

MANPOWER ALLOCATION PROBLEM WITH JOB-TEAMING CONSTRAINTS

4.1 Introduction

The Manpower Allocation Problem with Job-Teaming Constraints (MAPTC) is the problem of assigning workers to tasks, where each task requires a particular subset of workers to be present simultaneously in order to be accomplished. The objective is to minimize the total completion time of all tasks. Tasks and workers are heterogeneous, where the particular capabilities to accomplish tasks are distributed through out the workers; a given capability is unique to a single worker and tasks require different combination of capabilities to be accomplished.

Our motivating problem is a special case of single-task multi-robot instantaneous assignment (ST-MR-IA) problems with heterogeneous tasks and robots, in which each task requires a subset of robots to be present simultaneously in order to be accomplished. Multi-robot systems are becoming popular in robotics since multiple robots can solve tasks that single robots cannot solve or can solve them faster, with higher quality or in a more fault-tolerant way. There are different kinds of multi-robot task allocation problems (Gerkey and Mataric [24]): (i) robots might be capable of executing single or multiple tasks at the same time (single-task versus multi-task robots), (ii) tasks might require single or multiple robots for completion (single-robot versus multi-robot tasks), and (iii) the allocation of tasks to robots might be instantaneous or time-extended (instantaneous versus time-extended allocation).

These multi-robot task allocation problems consist of a number of targets, such as rocks on Mars from which one needs to take rock probes. Each target needs to be visited by one subset of robots simultaneously to perform some activity, which is the task associated with the target. Thus, robots arriving early at a target have to wait for the other robots to arrive. We propose an integer programming formulation and, since the problem is NP-hard, we also develop heuristic algorithms to find "good" solutions with a smaller runtime. Our research thus complements research on ST-MR-IA problems by introducing a special case of the problem in which simultaneous presence of a subset of robots is required and by offering solution approaches to this problem. This special case was previously studied by Parker and Tang [42], Zheng and Koenig ([60], [61]) in form of multi-robot routing problems with heterogeneous tasks and robots.

The simultaneous presence of workers for task accomplishment combined with objective function of minimizing the total completion time of all tasks is the novel part of our problem. In literature, manpower allocation problems with time windows (MAPTW) which require synchronization of servicemen for service completion has previously treated by Bredstrom and Ronnqvist [6], Dohn et al. [13] and Li et al. [34]. Lim et al. [35] study a manpower allocation problem motivated by the Port of Singapore and formulate it as a multi-objective problem with the primary objective of minimizing the number of servicemen used while satisfying all service demands. They do not require simultaneous work of servicemen whereas each demand location requires a given number of servicemen to perform work within its time window. Instead, they can start serving the same demand location at different times as long as they start after the early time and finish before late time, which define a time window for that particular location. Lim et al. [35] present meta-heuristic approaches for the problem and compare them on randomly generated problem instances. Li et al. [34] consider the same problem with the addition of job-teaming constraints, i.e., all teammates have to work simultaneously on a given task in a given time window, which is defined as the manpower allocation problem with time windows and job-teaming constraints (MAPTWTC). The objective function is a weighted sum of the total number of workers and the total travel time of all workers. Li et al. [34] develop construction heuristics with simulated annealing and test them on two types of generated cases (low and high ratio, where ratio is the average job duration to time window ratio). Lower bounds for the number of required workers are obtained from a network flow model and algorithms are compared to optimal solutions in the high ratio case or lower bounds in the low ratio case. Bredstrom and Ronnqvist [6] study a vehicle routing and scheduling problem with time windows which is enforced to handle synchronization constraints. The problem is formulated as a multiple traveling salesman problem (TSP) and then temporal constraints enforcing synchronization and precedence are added. Dohn et al. [13] study m-MAPTWTC, where there are m teams to be assigned to tasks and the objective function is to maximize the number of assigned tasks. This study focuses on the scheduling of ground handling tasks in some of Europe's major airports. The contribution of this research is the synchronization between teams in an exact optimization context. A branch-and-price approach is used to solve the problem optimally.

Our problem could be decomposed into subproblems, each corresponding to one worker if there were no teaming constraints. Teaming constraints enforce simultaneous work of workers on a task, i.e., arrival of a worker to task is not sufficient for processing of a task to start, all required workers should arrive to start processing. Solving the problem for each worker independently will give arrival times to each task, however we need to determine the start time of each task by choosing the largest among arrival times of all required workers for that task. Then, the start time of task will determine the arrival time to the next task for required workers. Tasks need to be visited simultaneously by several workers, therefore swapping two tasks in the visitation sequence of a worker does not only affect the swapped tasks and the worker, but also all required workers by these tasks and the following tasks in sequence for these workers. Furthermore, the objective function considered in our problem is totally different from the one of MAPTWTC. In our case, there are no time windows, corresponding to time windows going from zero to infinity, which corresponds to the smallest possible ratio of the average job duration to time window size in Li et al. [34]. For low ratio problems, they are not able to solve MAPTWTC optimally and only develop heuristics for their objective function, namely minimizing the number of workers or maximizing the number of assigned tasks within the given time windows. The objective function is to minimize the total task completion time of all tasks. We cannot eliminate any possible order of tasks (due to the missing time windows) and the search space is thus the largest possible.

The remainder of the chapter is organized as follows. In Section 4.2, we provide a formal definition of the problem and discuss its complexity. In Section 4.3, we present the MIP formulation. We explain the heuristic algorithms in Section 4.4. In Section 4.5, we provide the results of the computational experiments. Finally, we conclude with future directions in Section 4.6.

4.2 Problem Definition and Complexity

In this section, we define MAPTC formally and analyze its complexity. A set of workers, with specified initial locations, move on a connected graph $G = (V, E)$. $E \subseteq V \times V$ is the set of edges where each edge $(i, j) \in E$ has a given distance $d_{ij} \geq 0$. The set of nodes is $V = R \cup X$, where R denotes the nonempty set of initial nodes of the workers and X denotes the nonempty set of tasks. Thus, a node refers to either the initial location of a worker or the location of a task. With a slight abuse of notation, we use R both for the set of nodes for the initial locations of the workers and the set of workers. Let X^r be the nonempty set of tasks that require worker r for $r \in R$. We assume that the graph is complete since all connected graphs can be transformed into complete graphs using the standard Floyd-Warshall algorithm by computing the shortest paths in $O(|V|^3)$ time.

Each worker $r \in R$ starts at its initial node at time zero and then moves from a node to another node along the edges of the graph. In order for a task to be performed, a given subset of workers must be present at the location of the task. For task j, the subset of workers required is R^j. When all workers in R^j are not present in the location of the task j, workers who arrived early have to wait until the remaining workers arrive at the task. The objective is to minimize the total completion time, i.e. the latest start time among all tasks. We assume that the travel times are equal to the distances between nodes and that, for the completion of a task j, it is sufficient for all workers in R^j to visit task j at the same time, i.e. processing times are encoded as part of travel times. Our solution approach can also handle workers with different speeds by defining d_{rij} instead of d_{ij} without loss of generality.

It is trivial to show that solving MAPTC problem is NP-hard. Finding the minimum total distance for Hamiltonian path problems (HPPs) with one fixed endpoint is a special case of MAPTC where there is a single worker that has to start from its initial location and visit all tasks. Since there is a single worker, the objective of minimizing the total completion time simply equals minimizing the total distance. Hence, MAPTC reduces to finding the minimum total distance for HPPs starting at the initial location of the worker, and HPP with one fixed endpoint is NP-hard (Monnot [40]).

4.3 Mixed Integer Programming Formulation

In this section, we present the MIP formulation of the problem. The MIP formulation is motivated by the formulation of TSPs with the addition of start times of tasks. We also tried a sequence-based formulation which is different from typical formulations in the literature. This formulation does not have computational results as good as the first one, hence it is presented in Appendix C.

The decision variables are (i) the assignment variable x_{rij} which is 1 if worker r visits node i just before node j and 0 otherwise, (ii) the start time t_j of task j, which is the time task j is accomplished, (iii) the total completion time T of all tasks. For ease of formulation, we assume that each worker starts at its initial location and returns to its initial location, but we do not include the travel time of the return trip in the calculation of the total completion time in order to keep the setting of the problem unchanged. Using these decision variables, the problem is formulated as follows:

$$
\begin{aligned}
\textit{Minimize} \quad & T && (28)\\
\textit{subject to} \quad & && \\
\sum_{j \in X^r \cup \{r\} - \{i\}} x_{rij} &= 1 \quad i \in X^r \cup \{r\}, r \in R && (29)\\
\sum_{i \in X^r \cup \{r\} - \{j\}} x_{rij} &= 1 \quad j \in X^r \cup \{r\}, r \in R && (30)\\
t_j &\geq t_i + d_{ij} - M(1 - x_{rij}) \quad i \in X^r \cup \{r\}, j \in X^r, r \in R && (31)\\
T &\geq t_j \quad j \in X && (32)\\
x_{rij} &\in \{0,1\} \quad i \in X^r \cup \{r\}, j \in X^r \cup \{r\}, r \in R && (33)\\
t_i &\geq 0 \quad i \in R \cup X && (34)\\
T &\geq 0 && (35)
\end{aligned}
$$

In this formulation, (28) is the objective function which is the minimization of total completion time of all tasks. Constraints (29) ensure that workers visit one of the assigned tasks (or the worker's initial location) after visiting any one of the other assigned tasks. Constraints (30) guarantee that workers visit one of the assigned tasks before visiting any one of the other assigned tasks. Constraints (31) calculate the start times of the tasks by adding up the start time of the previously visited task and the distance from it to the next task. Constraints (32) states that the total completion time of all tasks must be greater than the start times of all tasks. Constraints (33), (34), and (35) are integrality and non-negativity constraints.

This formulation has the major drawback that the LP relaxation turns out to have an optimal solution with all t_j values and consequently T having values of zero. This stems from constraint (4), which does not force start times to increase for consecutive tasks in case of non-integral x_{rij} values. Therefore, we need to add some cuts to make this formulation stronger. We add constraints (36) to make sure that start times are not all 0. We also define a new set of variables, z_{rij}, which determine the arrival time of worker r to task j if task j is visited directly after task i by worker r to track the arrival times of workers individually. If task j is not visited directly after task i by worker r, then the z_{rij} value is identical to the start time of task i. Hence, instead of constraint (31), we add the constraints given below, say "Valid Inequalites-1 (VI-1)":

$$t_j \geq \sum_{i \in X^r \cup \{r\}} d_{ij} x_{rij} \quad j \in X^r, r \in R \tag{36}$$

$$z_{rij} = t_i + d_{ij} x_{rij} \quad i \in X^r \cup \{r\}, j \in X^r, r \in R \tag{37}$$

$$t_j \geq z_{rij} - M(1 - x_{rij}) \quad i \in X^r \cup \{r\}, j \in X^r, r \in R \tag{38}$$

Second set of valid inequalities can be generated by taking the total length of the path for each worker into consideration. We introduce the flow variables, y_{rij} (= the flow of worker r from node i directly to node j) is the amount of flow worker r carries from task i to task j. Then, "Valid Inequalites-2 (VI-2)" are:

$$\sum_{j \in X^r} y_{rjr} = |X^r| + 1 \quad r \in R \tag{39}$$

$$\sum_{i \in X^r \cup \{r\} - \{k\}} y_{rik} - \sum_{j \in X^r \cup \{r\} - \{k\}} y_{rkj} = 1 \quad k \in X^r, r \in R \tag{40}$$

$$(|X^r| + 1) x_{rij} \geq y_{rij} \quad i, j \in X^r \cup \{r\}, r \in R \tag{41}$$

$$y_{rij} \geq 0 \quad i, j \in X^r \cup \{r\}, r \in R \tag{42}$$

VI-2 consists of flow balance constraints (39), (40) and (41) and nonnegativity constraints for the flow variables (42). Constraints (39) state that the total flow of worker r is the total number of assigned tasks plus one (for returning to initial location of worker). Constraints (40) ensure that if worker r visits task k, it leaves one unit of flow at this node and then continues with the remaining amount. Constraints (41) guarantee that there is no flow if worker r does not visit task j after visiting task i.

Moreover, a lower bound for objective function T can be generated by Hamiltonian path:

$$T \geq \max_r HPbound_r \tag{43}$$

where $HPbound_r$ is generated by solving the HPP with one fixed endpoint for each worker r in which individual paths are generated for each worker starting at initial location of worker and visiting all tasks that requires worker r. By taking the maximum over all these paths, we generate a lower bound for total completion time. The HPP with one fixed endpoint for each worker can be solved since the problem size is $|X^r|$ for each worker r. However original problem is combination of these problems for all workers with synchronization constraints which makes decomposition impossible.

The objective function to be minimized does not include decision variables other than T explicitly, it is just enforced to be greater than or equal to all start times, t_j's, and in calculation of start times, x_{rij} variables are used in constraints (31) and (32). If objective was minimizing total distance traveled as in (44), decision variables x_{rij} would be explicitly used in the objective function.

$$Minimize \quad \textstyle\sum_{r\in R} \sum_{i\in X\cup\{r\}} \sum_{j\in X} d_{ij}x_{rij} \tag{44}$$

Minimizing (44) can be solved optimally for small instances. However, this will relax our synchronization assumption since t_j and T will not be included in minimization. Thus, we can try a combination of these two objectives, namely total completion time and total distance traveled as in (45) which will enforce the sum to be minimized and also include x_{rij} variables in the objective.

$$Minimize \;\; T + \textstyle\sum_{r\in R} \sum_{i\in X\cup\{r\}} \sum_{j\in X} d_{ij}x_{rij} \tag{45}$$

4.4 Heuristic Ideas

In this section, we propose heuristic ideas to solve the problem in smaller runtime and examine how the total completion time differs for various approaches. Total completion time includes the waiting time of early arriving workers to a task as well as total travel time. The waiting time at a task j is the difference between arrival times of earliest arriving worker and the latest arriving worker. The waiting time, the number of required workers - $|R^j|$ for task j - and the earliest possible start time of task j can be listed as significant factors affecting total completion time and can be used in heuristic approaches. We assume that a task j is **easier** than task i if $|R^j| < |R^i|$, i.e., task j requires less number of workers than task i; and **harder** otherwise. The proposed heuristic algorithms are listed as follows:

1. Easiest First (EF): Choose the task that requires the least number of workers among all unassigned tasks and send the required workers to that task. In case of equal number of workers required, choose the task that can be visited earlier. Continue until all tasks are visited.

2. Hardest First (HF): Choose the task that requires the most number of workers among all unassigned tasks and allocate the required workers to that task. In case of equal number of workers required, choose the task that can be visited earlier. Continue until all tasks are visited.

3. Earliest Start Time (EST) : Choose the task that can be started earliest among all other unassigned tasks.

4. Score-Based Heuristic (SB): Calculate scores for every unassigned task at every step by summing the earliest possible start time of the task and maximum of the waiting times of required workers. Earliest possible start time of the task can be defined as the maximum of arrival times of all required workers. Waiting

time of a worker can be defined as the difference between earliest possible start time and arrival time of the worker. Instead of choosing tasks according to earliest possible start time greedily, maximum waiting time is added to take into account the start times of remaining tasks.

Choose the task with smallest score and allocate the required workers to that task. Repeat same procedure until all tasks are visited. This approach tries to incorporate the waiting times while choosing the task to be visited at each step.

5. Sequence Bidding (SeqB): All workers bid a sequence for tasks that they are required to visit, i.e., each worker determines its preference for the sequence of tasks and assigns numbers from 1 to $|X^r|$ to tasks where numbers denote the preferred order of task in sequence of that worker. These bids are normalized by number of workers required - not to make the task which requires less workers more attractive. Hence, we can define bid for any task as summation of the order of task in the preferred sequence for every required worker, normalized by the number of workers required. Then, task with the lowest bid is chosen to be visited at each step. This procedure is repeated until all tasks are visited.

   The preferred sequence of tasks by each worker can be constructed using several methods, we use the Minimum Spanning Tree (MST) heuristic for constructing this sequence.

4.4.1 Improvements for Heuristics

After finding a solution by heuristics, i.e., finding a sequence of tasks to be visited for each worker; we exchange the tasks in the sequence if the total completion time decreases. In order to determine a single sequence including all tasks, the tasks are sequenced starting with the one which has the earliest start time to the one which is started latest according to the solution generated by heuristics. Then, we consider 2-exchanges on this sequence. Forward-Improvement (FI) starts with the first task (in sequence) and considers exchanging the first task with the second one and calculates the resulting total completion time T' with new sequence and performs the exchange if $T' \leq T$. Then, considers exchanging the second and third task in updated sequence, third and fourth task,... till the end of the sequence while performing 2-changes if T improves. Backward-Improvement (BI) starts with the last task (in sequence) and considers exchanging the last task with the $(|X| - 1)$ st one and again calculates the resulting total completion time T'' with new sequence and performs the exchange if $T'' \leq T$. Instead of considering only 2-exchanges with the nearest neighbor, we expand the neighborhood region and consider exchanging task k and $k + i$ for all $i \in \{1, 2, 3, 4, 5\}$ and for all tasks k in FI. In BI, we again consider exchanging task k and $k - i$ for all $i \in \{1, 2, 3, 4, 5\}$ and for all tasks k. BI and FI improvement steps are applied with varying neighborhood regions and varying orders (FI-BI: first FI, then BI and BI-FI:first BI, then FI). Improvement algorithm for FI-BI is presented in Algorithm 1, BI-FI is the same except for the order of FI and BI.

Algorithm 1 *Improvement Algorithm*

Let S_0 be the initial sequence of tasks according to start times by one of the heuristics, and $T(S_0)$ is the total completion time of sequence S_0
Forward Improvement where n is the neighborhood size
for $j = n$ to 1 **do**
  for $k = 1$ to $|X| - j$ **do**
    Construct S_1 by swapping tasks in sequence k and $(k + j)$
    if $T(S_0) \geq T(S_1)$ **then**
      Construct new S_0 by swapping the tasks in sequences k and $(k + j)$
    end if
  end for
end for
Backward Improvement where n is the neighborhood size
for $j = n$ to 1 **do**
  for $k = |X|$ to $n + 1$ **do**
    Construct S_2 by swapping tasks in sequence k and $(k - j)$
    if $T(S_0) \geq T(S_2)$ **then**
      Swap tasks in sequences k and $(k - j)$
    end if
  end for
end for

4.4.2 Improvements for Sequence Bidding Heuristic

In sequence bidding heuristic, each worker determines the preferred sequence of tasks and assigns numbers from 1 to $|X^r|$ to tasks. These sequences are generated by MST heuristic as explained in Johnson and Papadimitriou [28] which is a 2-approximate algorithm for TSP when distance matrix satisfies the triangle inequality. In this heuristic, a minimum spanning tree S is computed first, then an Eulerian graph is obtained by doubling edges of S and finally Hamiltonian cycle is obtained by shortcutting the tour. A tour starting and ending at root node r is obtained by applying the MST heuristic on a graph which has worker r as root node and the tasks that worker r is required to visit as other nodes. Let us define $taskSequence_{ri}$ $r \in R^i$, $i \in X$ which gives the order of task i in sequence of worker r generated by MST heuristic. Then, $taskScore_i$ is calculated which gives the normalized order for task i:

$$taskScore_i = \frac{\sum_{r \in R^i} taskSequence_{ri}}{|R^i|}, \forall i \in X \tag{46}$$

We also try using $taskSequence_{ri}/|X^r|$ instead of $taskSequence_{ri}$ in order to take into account relative position of a given task with respect to the total number of tasks that worker r has to visit:

$$taskScore_i = \frac{\sum_{r \in R^i} \left(\frac{taskSequence_{ri}}{|X^r|}\right)}{|R^i|}, \forall i \in X \tag{47}$$

However, the preferred sequence of tasks for all workers may not be equally important since some of these will have slack times (or waiting times for early arriving workers). Therefore, we calculate the total completion time of sequence generated by each worker and instead of using all these sequences, take only the ones with the largest total completion times. Sort the workers according to decreasing order of total completion time of their preferred sequences and define $n$umCriticalPaths which determines how many of these sequences will be used in heuristic. $n$umCriticalPaths is varied from 1 to (number of workers - 1), i.e. if $n$umCriticalPaths is 1, only the preferred sequence of worker with largest total completion time is used whereas if $n$umCriticalPaths is (number of workers - 1) then all but the preferred sequence of worker with smallest total completion time is used in $taskScore_i$ calculation. By this way, we are not taking the preferred sequence of workers that are not critical, i.e., not long as the *numCriticalPaths* longest ones. This algorithm is explained below:

Critical Sequence Bidding and Insertion Heuristic

1. Preferred Sequence Step for all workers: Calculate preferred sequence for each worker by MST heuristic.

2. Critical Sequence Determination Step: Choose the *numCriticalPaths* longest paths, and apply original Sequence Bidding algorithm for these paths using the following task scores. R_c^i denotes the workers required to visit task i which have critical paths.

$$taskScore_i = \frac{\sum_{r \in R_c^i} \left(\frac{taskSequence_{ri}}{|X^r|} \right)}{|R_c^i|}, \forall i \in X \tag{48}$$

3. Insertion Step: The resulting sequence of tasks may be a partial sequence, then add the tasks that are not currently in the sequence just by trying to insert every possible position and choose the position which results in the lowest total completion time.

4. Improvement Step: Apply improvement steps, BI (backward improvement) and FI (forward improvement) with varying neighborhood regions.

In the insertion step, while adding the tasks that are not currently in the sequence, we determine the order that we insert them to sequence either by their indices or probabilistically. Then, we apply the improvement step and report the one with minimum total completion time.

4.5 Computational Results

In this section, we compare the performance of heuristics and the MIP formulation. We have small instances (5 workers 25 tasks and 5 workers 50 tasks); and large instances (10 workers 50 tasks and 10 workers 100 tasks). In all these settings, the tasks and the workers are randomly located to the entire area of size 100x100 as in Ekici et al. [18]. The distances between the tasks and the workers are calculated as Euclidean distance.

Worker requirements of tasks are determined according to Algorithm (2).

Algorithm 2 *Target Data Generation*

1: Let $[a, b]$ define the range for number of workers required by any task and define $length$ as $(b - a + 1)$.
Let $workerAllocation_{rj}$ be a binary variable which takes value 1 if worker r is required by task j, and 0 otherwise. Initially all $workerAllocation_{rj}$ variables are 0.
2: **for** $j = 1$ to $|X|$ **do**
3: $\quad prob$ = random variable in $[0, 1)$
4: $\quad |R^j| = a + \lfloor (length * prob) \rfloor$
5: $\quad$ **for** $k = 1$ to $|R^j|$ **do**
6: $\quad\quad$ Choose among workers for which $workerAllocation_{rj} = 0$ randomly, say worker r' is chosen
7: $\quad\quad workerAllocation_{r'j} = 1$
8: $\quad$ **end for**
9: **end for**

For each setting, the range for number of workers required is varied and 2 scenarios, "Low Coordination" and "High Coordination" -according to amount of coordination tasks need- are constructed as in Table 8:

Table 8: Scenarios according to worker requirements of tasks.

	5 workers	10 workers
Low Coordination	[1,2]	[1,3]
High Coordination	[1,4]	[3,7]

For each setting and coordination level, 10 instances are generated and the average of total completion time for these 10 instances are reported in Table 9. "Setting" column consists of two numbers, first one is the number of workers, second one is number of tasks and "low", "high" represents how the worker requirement data is generated as given in Table 8. Average of 10 instances for all heuristics are given in columns: no (no improvement), FI-BI (first FI then BI applied with varying neighborhood regions), BI-FI (first BI then FI applied with varying neighborhood regions), min (minimum of FI-BI and BI-FI taken for each instance). Results are given for all heuristics: EF, HF, EST, SB, SeqBByIndex, SeqBRand where SeqBByIndex and SeqBRand are variants of critical sequence bidding heuristic with improvement steps in which insertion is done according to indices of tasks or probabilistically for the tasks not inserted earlier. The minimum of average total completion times are given in bold for each row. In Table 10, run-time of heuristics are given for each setting and coordination level. Except for large instances solved by SeqB heuristic, the run-times are less than 70 seconds. For large instances solved by SeqB, run-times increase up to around 3000 seconds (=1 hr).

Table 9: Heuristic results.

setting	coord	EF no	EF FI-BI	EF BI-FI	EF min	HF no	HF FI-BI	HF BI-FI	HF min	EST no	EST FI-BI	EST BI-FI	EST min
5_25	low	456.92	371.18	376.99	368.54	482.75	375.30	371.98	362.48	412.88	354.30	356.08	353.71
	high	671.15	504.29	500.25	490.15	706.29	514.83	509.99	507.68	500.16	468.42	470.30	468.42
5_50	low	667.34	564.44	566.18	560.38	736.30	577.61	562.42	558.63	633.62	541.91	533.17	**529.47**
	high	1018.28	807.33	791.55	782.50	1067.89	811.09	829.21	799.04	737.63	683.12	675.48	674.99
10_50	low	742.37	554.61	547.74	541.95	772.35	554.09	547.18	537.07	626.87	543.61	543.60	534.53
	high	1413.42	936.62	934.16	921.53	1387.62	913.70	901.26	883.19	720.21	694.69	696.68	694.02
10_100	low	1135.10	930.82	916.60	908.50	1165.46	930.70	929.39	918.78	923.95	840.74	834.61	**828.91**
	high	2037.49	1447.88	1488.26	1437.69	1995.52	1458.48	1412.06	1391.66	981.41	938.53	934.01	933.45
All	low	750.43	605.26	601.87	594.84	789.22	609.42	602.74	594.24	649.33	570.14	566.86	561.66
	high	1285.08	924.03	928.55	907.97	1289.33	924.52	913.13	895.39	734.85	696.19	694.12	692.72
Overall		1017.76	764.65	765.21	751.40	1039.27	766.97	757.94	744.82	692.09	633.16	630.49	627.19

setting	coord	SB no	SB FI-BI	SB BI-FI	SB min	SeqBByIndex no	SeqBByIndex FI-BI	SeqBByIndex BI-FI	SeqBByIndex min	SeqBRand no	SeqBRand FI-BI	SeqBRand BI-FI	SeqBRand min
5_25	low	409.87	361.48	361.47	359.39	444.02	356.19	353.09	350.87	435.94	351.60	351.56	**345.07**
	high	528.24	460.95	463.39	460.78	648.84	454.18	451.12	**440.75**	623.15	451.84	461.71	443.48
5_50	low	616.32	557.34	552.20	549.61	794.15	545.06	556.13	538.40	760.73	544.31	552.97	537.91
	high	725.52	682.43	676.67	**674.59**	1098.56	749.07	744.56	726.75	1055.90	738.16	741.88	728.84
10_50	low	627.76	550.16	552.01	540.37	740.57	516.57	515.26	506.13	695.78	511.39	510.72	**498.33**
	high	727.80	693.50	687.94	**685.90**	1219.64	765.81	766.81	747.91	1195.43	772.41	771.26	753.96
10_100	low	944.05	844.91	855.96	842.55	1445.60	913.77	899.69	895.08	1419.59	899.44	903.96	885.31
	high	975.84	927.02	927.14	**925.26**	2081.80	1259.48	1256.21	1244.48	2030.42	1204.50	1242.25	1197.83
All	low	649.50	578.47	580.41	572.98	856.08	582.90	581.04	572.62	828.01	576.68	579.80	566.65
	high	739.35	690.98	688.78	686.63	1262.21	807.14	804.67	789.97	1226.22	791.73	804.27	781.03
Overall		694.43	634.72	634.60	629.81	1059.15	695.02	692.86	681.30	1027.12	684.21	692.04	673.84

Table 10: Heuristic run-time in seconds.

setting	coordination	EF	HF	EST	SB	SeqBByIndex	SeqBRand
5_25	low	0.9	0.6	1.1	0.3	6.2	6.3
	high	1.2	0.8	0.8	0.5	7.6	5.8
5_50	low	4.7	4.5	5.4	3.5	40	53.9
	high	7.9	6.1	7.4	4.8	50.6	67.7
10_50	low	5.3	6.7	4.9	5.5	700.7	234
	high	9.2	9.7	8.2	7.2	465.1	343.6
10_100	low	35.9	40.4	33.5	35.9	1677.8	3070.4
	high	60.6	69.3	58.9	59.9	2762.7	2961.9

Percent change in total completion time after improvement steps (assuming post processing is performed and minimum of FI-BI, BI-FI is chosen) for each setting and coordination level is given in Table 11. Overall, for all instances and all heuristics 22.5% improvement is obtained by improvement steps. The largest improvement is obtained in SeqB heuristics (33.24% and 31.45% respectively for SeqBByIndex and SeqBRand) and the least in EST and SB (9.52% and 9.40%). For low coordination instances, 21.46% improvement is obtained whereas for high coordination instances, 23.54% improvement is obtained on average.

Paired t-test is used to compare heuristics pairwise and to conclude which heuristic performs better in each setting and coordination level. The differences in total completion time for every pair of heuristics are calculated for each instance. Then paired t-test is performed as in Walpole et al. [54] to conclude one is better than the other for given setting and coordination level. We cannot conclude one heuristic is better than the others. This is an expected result since heuristics are similar in their criteria. Therefore, we group heuristics according to their criteria: 1- (EF, HF): number of workers required for each target is the main criteria in both heuristics, 2- (EST, SB): earliest start time is the main criteria, 3-SeqB includes SeqBByIndex, SeqBrand; and take the average within group for each instance. Then we conclude that (EST,SB) heuristics perform better than the others. Table 12 displays the results of the paired t-test for all instances (including all settings and coordination levels)

Table 11: Percent improvements for heuristics.

setting	coord	EF	HF	EST	SB	SeqBBy Index	SeqB Rand	average
5_25	low	18.41%	24.55%	12.83%	11.19%	20.49%	20.45%	17.99%
	high	26.75%	27.59%	6.15%	11.99%	31.90%	28.52%	22.15%
5_50	low	15.15%	23.56%	16.29%	10.75%	32.06%	29.17%	21.16%
	high	22.86%	25.06%	8.10%	6.60%	33.75%	30.45%	21.14%
10_50	low	26.63%	29.48%	14.11%	13.59%	31.34%	28.13%	23.88%
	high	34.71%	36.27%	3.57%	5.55%	38.44%	36.47%	25.83%
10_100	low	19.54%	21.06%	10.33%	10.52%	37.83%	37.48%	22.79%
	high	29.35%	29.95%	4.81%	5.03%	40.13%	40.96%	25.04%
All	low	19.93%	24.66%	13.39%	11.51%	30.43%	28.81%	21.46%
	high	28.42%	29.72%	5.66%	7.29%	36.05%	34.10%	23.54%
Overall		24.17%	27.19%	9.52%	9.40%	33.24%	31.45%	22.50%

where the p-values smaller than 0.10 are given in bold. The differences are calculated by subtracting the total completion time of heuristic in the lower level from that of heuristic in the upper level in first row. Comparing EF and HF, the mean of differences is 7.28 (the average difference of total completion times calculated by EF are 7.28 more than that of HF) and p-value is 0.29, hence we cannot conclude that these two heuristics are statistically different. Same applies for SB - EST and SeqBByIndex - SeqBRand comparisons with p-values respectively 0.42 and 0.81. Then, groups of heuristics are compared and according to results, we conclude that (EST,SB) is better than (EF,HF) and SeqB is better than (EF,HF), and (EST,SB) is better than SeqB with p-values=0.00. Therefore, we can rank groups as follows from the best to the worst according to the objective values: (EST,SB), SeqB, (EF,HF).

Heuristic comparisons are also performed for each setting and coordination level given in Table 13. Differences are again calculated by subtracting the total completion time of heuristic in the lower level from that of heuristic in the upper level in the first row; and if this difference is negative, p-value is given in italic. For 5_25, low, high and all coordination instances, SeqB is significantly better than other heuristics. For 5_50, high and all coordination instances (EST,SB) is significantly better than others,

Table 12: Paired t-test for Heuristics.

	EF HF	SB EST	SeqBByIndex SeqBRandom	(EF,HF) (EST,SB)	(EF,HF) SeqB	SeqB (EST,SB)
mean	7.28	4.11	0.82	119.61	70.54	49.07
std dev	61.00	45.72	30.31	159.68	71.80	105.48
t statistic	1.07	0.80	0.24	6.70	8.79	4.16
p-value	0.29	0.42	0.81	**0.00**	**0.00**	**0.00**

Table 13: Paired t-test for heuristics for each setting and coordination level.

setting	coordination	EF HF	SB EST	SeqBByIndex SeqBRandom	(EF,HF) (EST,SB)	(EF,HF) SeqB	SeqB (EST,SB)
5_25	low	0.49	0.27	0.83	0.12	**0.01**	*0.05*
	high	*0.28*	*0.32*	*0.12*	**0.05**	**0.00**	*0.05*
	all	*0.67*	*0.86*	*0.35*	**0.02**	**0.00**	*0.01*
5_50	low	0.75	0.27	0.79	**0.09**	**0.09**	*0.90*
	high	*0.01*	0.94	0.77	**0.00**	**0.00**	**0.00**
	all	*0.09*	0.37	0.68	**0.00**	**0.00**	**0.02**
10_50	low	0.96	0.58	0.62	0.86	**0.00**	*0.00*
	high	0.15	*0.55*	*0.74*	**0.00**	**0.00**	**0.00**
	all	0.19	*0.99*	1.00	**0.00**	**0.00**	0.33
10_100	low	*0.56*	0.39	*0.65*	**0.00**	**0.04**	**0.00**
	high	**0.04**	*0.64*	0.27	**0.00**	**0.00**	**0.00**
	all	0.14	0.61	0.53	**0.00**	**0.00**	**0.00**

however for low coordination instances there is no significant difference between SeqB and (EST,SB). For 10_50, low coordination instances, SeqB is significantly better than others and for high coordination instances (EST,SB) is significantly better than others with p-values 0.00. For 10_100, low, high and all coordination instances, the ranking (EST,SB) SeqB, (EF,HF) is valid with p-values less than 0.04.

Another possible comparison can be made for improvement steps to analyze the effect of the order in which FI and BI are applied. For this purpose, we performed paired t-tests for all heuristics for values with FI-BI and BI-FI. However, we cannot conclude that they are significantly different with p-values > 0.05.

The results for MIP formulations are presented in Table 14. In all scenarios, $(VI-1)$ is used since without these valid inequalities all t_j values and therefore T turns out to be zero. The MIP formulations are varied in six ways to construct various scenarios: (i) (T) or $(T+)$: objective function T, or combined objective function $(T+)$ in Equation (45) (in each case, T values are reported in table), (ii)

$(VI{-}2)$: valid inequality set with flow variables are added or not, (iii)(HP): HP-bound for T added or not, Equation (43), (iv) (WS): specifies whether MIP is initialized by solutions with objective function values given in "WS initial solution" column (Warm-start(WS) initial solution is the solution with the minimum objective function found so far by heuristics) is used or not, (v) (M): the value of M in the constraints is originally calculated according to given distance data of the instance of the problem, but after WS we can replace this M value as objective function of WS initial solution which would be tighter than the original value, (vi) (B): specifies if different branching strategies such as branch-up and SOS variables are used or not. Run-time for scenarios without WS is 10 hours whereas for scenarios with WS run-time is 6 hours.

Table 14: MIP formulation results.

setting	coord	(T)	(T) (VI-2)	(T) (HP)	(T) (VI-2) (HP)	(T+) (HP)	(T+)(VI-2) (HP)	WS initial soln	(T) (HP) (WS)	(T)(HP) (WS) (M)	(T) (HP) (WS) (M) (B)	(T) (VI-2) (HP) (WS)	(T) (VI-2) (HP) (WS) (M)	(T) (VI-2) (HP) (WS) (B)	instance based min
5_25	low	331.68	380.79	376.22	378.77	352.09	348.89	328.14	324.87	325.62	325.48	327.29	326.31	325.54	324.87
	high	818.15	N/A	850.59	N/A	432.52	434.82	419.08	417.70	418.30	417.32	418.33	418.52	418.65	416.67
5_50	low	896.62	N/A	977.41	N/A	508.12	471.11	467.22	456.24	455.41	456.97	462.19	460.87	461.38	452.44
	high	1805.36	N/A	1925.77	N/A	1935.27	556.41	633.58	626.07	627.03	626.91	632.23	631.91	629.97	623.13
10_50	low	932.61	1178.48	890.47	1178.48	1272.23	488.05	487.55	479.65	478.64	474.33	480.29	481.95	481.23	472.22
	high	2410.44	N/A	2432.31	N/A	2546.16	595.01	669.29	668.98	669.29	669.29	669.29	669.29	669.29	668.98
10_100	low	2230.92	N/A	2378.96	N/A	2730.24	N/A	806.54	806.44	805.84	804.25	806.05	806.54	805.89	804.19
	high	5178.45	N/A	5178.58	N/A	5178.45	N/A	911.43	911.43	911.43	911.43	911.43	911.43	911.43	911.43

Table 15: Paired t-test for MIP Scenarios.

setting	coord	I (T) (T)(VI-2)	II (T) (T)(HP)	III (T)(HP) (T+)(HP)	IV (T)(VI-2)(HP) (T+)(VI-2)(HP)	V (T+)(HP) (T+)(VI-2)(HP)	VI (T+)(HP) (T)(HP)(WS)	VII (T)(HP)(WS) (T)(VI-2)(HP)(WS)	VIII (T)(HP)(WS) (T)(HP)(WS)(M)	IX (T)(HP)(WS) (T)(HP)(WS)(B)	X (T)(VI-2)(HP)(WS) (T)(VI-2)(HP)(WS)(M)	XI (T)(VI-2)(HP)(WS) (T)(VI-2)(HP)(WS)(B)
5_25	low	*0.00*	*0.01*	0.17	**0.10**	0.34	**0.01**	*0.12*	*0.34*	*0.34*	0.23	**0.08**
	high	N/A	*0.63*	**0.00**	N/A	*0.65*	**0.01**	*0.54*	*0.51*	0.21	*0.71*	*0.34*
5_50	low	N/A	*0.26*	**0.00**	N/A	0.17	**0.07**	*0.05*	0.74	*0.59*	0.65	0.59
	high	N/A	*0.14*	*0.95*	N/A	0.19	**0.00**	*0.03*	*0.67*	*0.68*	0.36	0.22
10_50	low	N/A	0.56	*0.00*	N/A	0.00	**0.00**	*0.84*	0.75	0.19	*0.33*	*0.71*
	high	N/A	*0.60*	*0.01*	N/A	N/A	**0.00**	*0.34*	*0.34*	*0.34*	N/A	*0.34*
10_100	low	N/A	*0.21*	*0.01*	N/A	N/A	**0.00**	0.42	0.69	**0.04**	*0.30*	0.41
	high	N/A	*0.35*	0.35	N/A	N/A	**0.00**	0.86	0.17	**0.08**	0.34	**0.10**

Table 16: Improvement compared to MIP after WS.

instance	coord	Heuristic	MIP before WS	WS initial soln	MIP after WS	% imp heuristic	% imp MIP before WS	% imp WS initial soln
5_25	low	341.16	331.24	328.14	324.87	4.64%	1.83%	0.96%
	high	429.43	430.50	419.08	416.67	2.83%	3.14%	0.57%
5_50	low	509.58	470.54	467.22	452.44	11.16%	3.87%	3.25%
	high	658.84	1396.18	633.58	623.13	5.53%	44.27%	1.63%
10_50	low	494.01	722.98	487.55	472.22	4.40%	27.05%	3.20%
	high	676.09	2069.20	669.29	668.98	1.07%	59.96%	0.05%
10_100	low	806.54	2192.97	806.54	804.19	0.30%	62.73%	0.30%
	high	911.43	5178.45	911.43	911.43	0.00%	82.13%	0.00%
overall	low	537.82	829.87	522.36	513.43	5.13%	20.54%	1.93%
	high	668.95	2120.87	658.35	655.05	2.36%	43.85%	0.56%

Several different scenarios are constructed by combinations of variations explained above, in order to understand the effect of each variation and results are presented in Table 14. Six different scenarios before WS and after WS are used for comparison. Before WS, objective function is varied: (T) or (T+), (VI-2) is added and (HP) bound is added. It should be noted that with addition of (VI-2), MIP cannot find any feasible solution in given run-time for large instances whereas it was able to find without (VI-2). (T+)(VI-2)(HP) scenario seems to find the lowest average values for all instances except for 10_100 instances for which it cannot find any feasible solution in 10 hours. Then, among all heuristic solutions found so far, the lowest for each instance is used as initial solution, called "WS initial solution". After WS, the objective function is used as T and (HP) is added to all scenarios since we are providing an initial solution which is above this bound. Then, we used M value as the initializing solution which will be the tightest possible with given solution and different branching strategies in order to improve the results of WS solution. For instance, we used *branch-up* strategy for binary decision variables x_{rij} since the 0-branch has a very little effect on the objective value whereas the 1-branch usually results in significant change in solution and objective values. In CPLEX, *branch-up* strategy is chosen which influences the direction, may be up or down - enforced up in this strategy, of the branch on that variable to be explored first after a variable has been selected for branching. Moreover, we deployed Special-Ordered-Sets (SOSs) to specify integrality conditions which restrict the number of nonzero solution values among a specified set of variables in the model. We used SOSs for x_{rij} variables among which only one of them will be nonzero for each robot r and each target i or j. In the last column overall instance-based minimum is given for each setting and coordination level.

Paired t-test is used to compare different scenarios given in Table 14. Table 15 displays the p-values for the comparison of T values for scenarios in the upper and lower levels of the first row for each setting and coordination level. If the mean of differences are negative between the upper and lower level, the difference between second and first row is used and p-value is given in italic (italic means scenario in the upper level gives better results). It is interesting to note that for 5_25 low setting, addition of (HP) and (VI-2) before WS -given in columns I and II- results in a significant decrease (p-values 0.00 and 0.01). The reason for better solutions for (T) scenario may be addition of (VI-2) restricts the solutions and it becomes harder to find any starting feasible solution - no feasible solution were found in (T)(VI-2) scenarios for settings except 5_25 low. Addition of HP bound may force to find a better initial feasible solution since it has to satisfy HPbound, however in most of the settings it does not lead to better final solutions although not significantly worse than those found in (T) scenario. Using the combined objective function (column III) results in significant decrease for small instances whereas for large instances it results in significant increase given in italic. This can be explained by nature of the combined objective, minimizing total distance traveled helps for smaller instances but not for larger instances. For all settings and coordination levels, WS improves significantly as given in column VI. Addition of M does not result in a significant change as can be seen in columns VIII and X, whereas branching results in significant for several settings - refer to columns IX and XI.

Table 16 displays the average objective function value for each setting within several steps. All the values given in Table 16 are calculated instance-based, i.e. minimum for each instance is taken and then averaged. Percent improvements are given for "heuristic", "MIP before WS" and "WS initial solution" with respect to "MIP after WS". "Heuristics" are improved by 5.13% for low coordination instances, whereas high coordination instances are improved by 2.36% on average. On the other hand, "MIP before WS" is improved by 20.54%, and 43.85% for low and high coordination instances, respectively. Note that the percent improvements on "MIP before WS" are very high for large instances, especially high coordination instances. In these cases, MIP was not able to find good solutions and starting MIP with heuristic solutions and solving again improved the results a lot. Slight improvements are obtained on WS solutions by running MIP again: 1.93% and 0.56% for low and high coordination levels, respectively.

4.6 Conclusion

We studied a manpower allocation problem with job-teaming constraints with the objective of minimizing total completion time of all tasks in which each worker has different functionalities and each task requires different combination of these functionalities. We developed a MIP formulation for this problem and showed the difficulty of objective function even for smaller instances. Then we defined set of valid inequalities and developed an easier combined objective function to improve the objective function value. Although the results for small instances improved with these enhancements, for larger instances the results need more improvement. We developed several heuristic approaches and used their best solutions to initialize MIP which improved the results for large instances. Furthermore, we used different branching strategies to improve the solutions initialized.

We evaluated the performance of heuristics and MIP formulation with different strategies on test data sets originating from multi-robot task allocation problems. Each robot has different functionalities, and each task requires different combinations of these functionalities. For example, taking rock probes requires drilling the rock, image processing to drill at the right position, keeping the rock steady and potentially cooling the drill. A robot with all of these functionalities would be large, heavy and expensive. It therefore makes sense to distribute the functionalities among robots (Lundh et al. [37]). As an extension of this problem, we can consider the problem in which we have sets of subsets, i.e., a target needs to get visited either by robots A and B or by robots B and C at the same time. This will complicate the problem since a task can be performed by different sets of robots and which set to choose among these sets should also be decided.

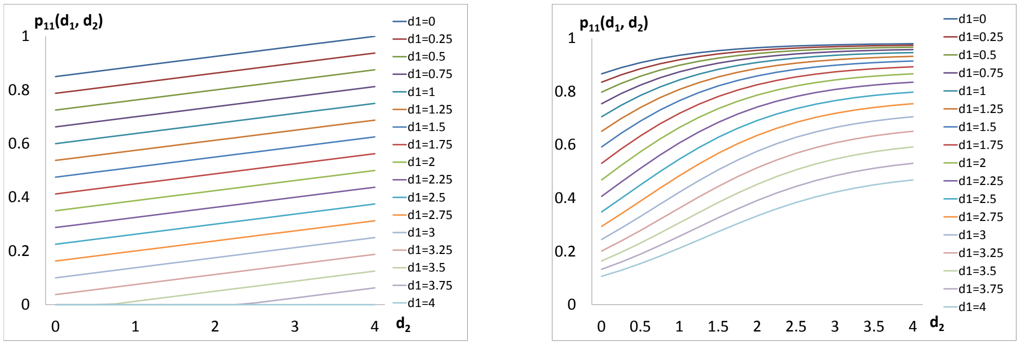
APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Comparison to MNL

MNL model is extensively used in marketing literature to model the choice behavior. van Ryzin and Mahajan [53], Maddah and Bish [38] and Suh and Aydin [52] are few examples considering assortment, inventory and pricing problems. van Ryzin and Mahajan [53] use a MNL model to describe the consumer choice process with corresponding preferences for each product and analyze the assortment variety inventory trade-off. Maddah and Bish [38] extend to include the pricing decisions. Another study using MNL and considering pricing decisions is Suh and Aydin [52] which use MNL model to calculate the purchase probabilities given utilities and prices of products. Comparing the MNL model according to quoted lead-times instead of quoted prices as in Suh and Aydin [52], we use lead-time and cross lead-time effects. Figure 21 shows a comparison of MNL and our choice model in the same lead-time quote space for choice probability of product 1 and 2, respectively for customer type 1. We use a linear model to be able to analyze the behavior of profit function in subsection 2.4.

The MNL model to determine the choice probabilities for each product and probability of not placing an order can be defined as follows. The choice set in our model consists of products 1, 2 and in addition, no order placement option which is denoted by index $i = 0$. Each option has an associated utility, u_{ci}, $i = 0, 1, 2$, $c = 1, 2$ for each customer type c where u_{c0} (utility of no order placement) can be assumed 0 without loss of generality. Given the options of products 1, 2 and no order placement option; customer chooses the option with largest utility. According to MNL model,



(a) $p_{11}(d_1, d_2)$ with $k_{12} = 0.15$ (b) MNL with $u_{11} = 4$ and $u_{12} = 2$

Figure 21: Choice probability for customer type 1 for product 1, preferred product.

the probability that option i has the largest utility for customer c is defined as follows:

$$f_{11}(d_1, d_2) = \frac{e^{u_{11}-d_1}}{1 + e^{u_{11}-d_1} + e^{u_{12}-d_2}}, \tag{49}$$

$$f_{12}(d_1, d_2) = \frac{e^{u_{12}-d_2}}{1 + e^{u_{11}-d_1} + e^{u_{12}-d_2}}, \tag{50}$$

$$f_{21}(d_1, d_2) = \frac{e^{u_{21}-d_1}}{1 + e^{u_{21}-d_1} + e^{u_{22}-d_2}}, \tag{51}$$

$$f_{22}(d_1, d_2) = \frac{e^{u_{22}-d_2}}{1 + e^{u_{21}-d_1} + e^{u_{22}-d_2}}, \tag{52}$$

where $f_{ci}(d_1, d_2)$ denotes the probability of placing an order of product i for customer type c.

Comparing the MNL model given above with our choice model, the implied utilities for any (d_1, d_2) pair will be as follows:

$$e^{\overline{u}_{11}-d_1} = \frac{p_{11}(d_1, d_2)}{p_{10}(d_1, d_2)} \quad or \ \overline{u}_{11} = ln\frac{p_{11}(d_1, d_2)}{p_{10}(d_1, d_2)} + d_1, \tag{53}$$

$$e^{\overline{u}_{12}-d_2} = \frac{p_{12}(d_1, d_2)}{p_{10}(d_1, d_2)} \quad or \ \overline{u}_{12} = ln\frac{p_{12}(d_1, d_2)}{p_{10}(d_1, d_2)} + d_2, \tag{54}$$

where $\overline{u}_{11}$ and $\overline{u}_{12}$ denotes the implied utility for (d_1, d_2) in our choice model corresponding to MNL model.

A.2 Optimality Equation for Finite State Space

For finite state space and finite lead-time quote space case, the optimality equation can be rewritten as follows, where lead-time quotes are incremented by 0.05 as explained in Section 2.5:

$$\frac{g^*}{\nu} + v^*(x_1, x_2) = \max_{d_1 \in (0, d_1^{\max}], d_2 \in (0, d_2^{\max}]} \Bigg\{ \Psi^1_{x_1,x_2}(d_1, d_2) + \Psi^2_{x_1,x_2}(d_1, d_2) + \left(\frac{\lambda_1}{\nu} p_{10}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{20}(d_1, d_2) \right) v^*(x_1, x_2) \Bigg\} \tag{55}$$

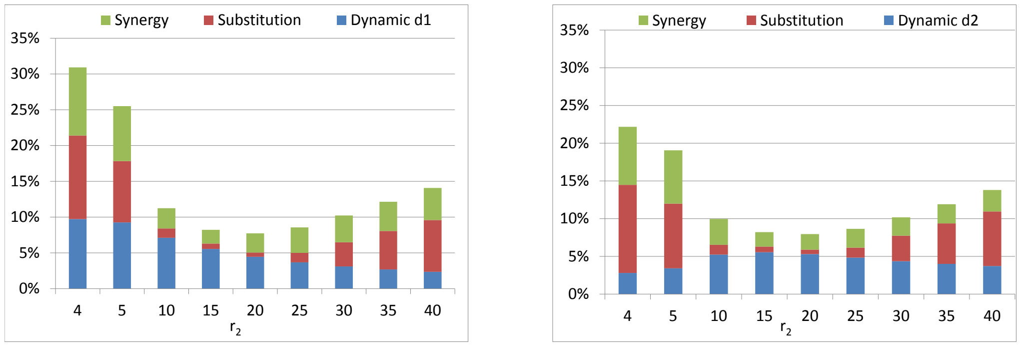
where $\Psi^1_{x_1,x_2}(d_1, d_2)$ and $\Psi^2_{x_1,x_2}(d_1, d_2)$ are defined as follows for state space $S = \{(x_1, x_2) : x_1 \in [0, N_1] \ and \ x_2 \in [0, N_2]\}$:

$$\Psi^1_{x_1,x_2}(d_1, d_2) = \begin{cases} \left(\frac{\lambda_1}{\nu} p_{11}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2) \right) (r_1 - l_1 L(x_1, d_1) + v^*(x_1 + 1, x_2)) + \\ \frac{\mu_1}{\nu} v^*(x_1, x_2), \text{for } x_1 = 0 \\ \left(\frac{\lambda_1}{\nu} p_{11}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2) \right) (r_1 - l_1 L(x_1, d_1) + v^*(x_1 + 1, x_2)) + \\ \frac{\mu_1}{\nu} v^*(x_1 - 1, x_2), \text{for } x_1 \in \{1, 2, ..., N_1 - 1\} \\ \left(\frac{\lambda_1}{\nu} p_{11}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{21}(d_1, d_2) \right) (r_1 - l_1 L(x_1, d_1) + v^*(x_1, x_2)) + \\ \frac{\mu_1}{\nu} v^*(x_1 - 1, x_2), \text{for } x_1 = N_1 \end{cases} \tag{56}$$

and

$$\Psi^2_{x_1,x_2}(d_1, d_2) = \begin{cases} \left(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2) \right) (r_2 - l_2 L(x_2, d_2) + v^*(x_1, x_2 + 1)) + \\ \frac{\mu_2}{\nu} v^*(x_1, x_2), \text{for } x_2 = 0 \\ \left(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2) \right) (r_2 - l_2 L(x_2, d_2) + v^*(x_1, x_2 + 1)) + \\ \frac{\mu_2}{\nu} v^*(x_1, x_2 - 1), \text{for } x_2 \in \{1, 2, ..., N_2 - 1\} \\ \left(\frac{\lambda_1}{\nu} p_{12}(d_1, d_2) + \frac{\lambda_2}{\nu} p_{22}(d_1, d_2) \right) (r_2 - l_2 L(x_2, d_2) + v^*(x_1, x_2)) + \\ \frac{\mu_2}{\nu} v^*(x_1, x_2 - 1), \text{for } x_2 = N_2 \end{cases} \tag{57}$$

A.3 Extra Computational Study



(a) Dynamic d_1 (b) Dynamic d_2

Figure 22: The impact of r_2 on benefits obtained by Dd_1 (Dd_2), substitution and synergy of both on NS-Sd1Sd2.

Figure 22 depicts the magnitudes of benefits obtained by substitution, dynamic quotation and the synergistic effects applied on base scenario, NS-Sd1Sd2. The benefits obtained by dynamic quotation of d_1 decreases as the revenue of other product increases, whereas the benefits obtained by dynamic quotation of d_2 increases as revenues of products get closer. On the other hand, the benefits obtained by substitution and the synergistic effects of dynamic quotation and substitution decrease as revenues of products get closer.

Figure 23 depicts the magnitudes of benefits obtained by substitution, dynamic quotation and the synergistic effects applied on base scenario, NS-Sd1Sd2. The benefits obtained by dynamic quotation of d_1 and d_2, and substitution decreases as proportion of customers get closer. On the other hand, the synergistic effects increase as proportion of customers get closer.

Observation 12 *As total traffic intensity increases, the benefit of dynamic quotation and/or substitution increases.*

Observation 12 can be seen from Figure 24. Dynamic quotation and/or substitution increases the profits more under higher traffic intensity. The improvements

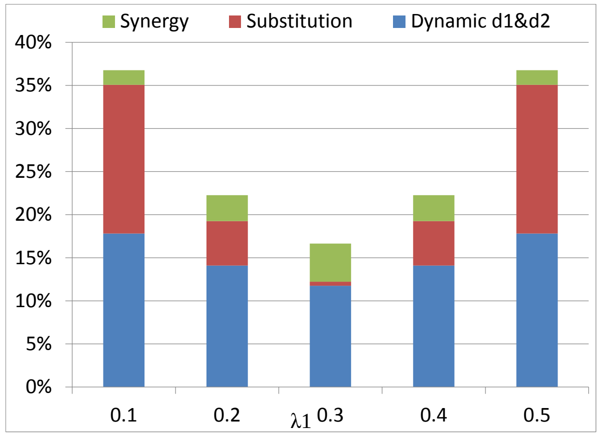


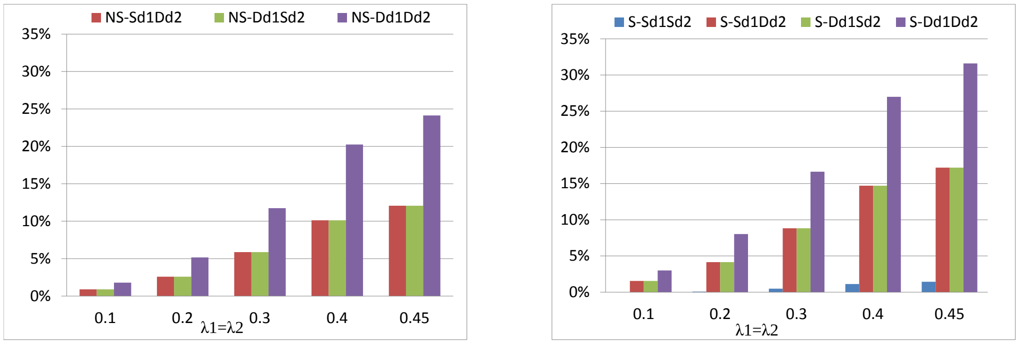
Figure 23: The impact of λ_1 on benefits obtained by Dd_1 (Dd_2), substitution and synergy of both on NS-Sd1Sd2 where $\lambda_1 + \lambda_2 = 0.6$.

are much higher when both products are quoted dynamically (NS-Dd1Dd2 and S-Dd1Dd2) compared to one of the products quoted static other quoted dynamic (NS-Sd1Dd2, NS-Dd1Sd2 and S-Sd1Dd2, S-Dd1Sd2). Note that Sd1Dd2 and Dd1Sd2 scenarios are identical since all parameters are equal for both products.

A.3.1 Impact of Lead-time Sensitivity

In this section, we test the impact of lead-time sensitivity on improvements on base scenario, NS-Sd1Sd2. The numerical analysis settings are summarized in Table 17. We test three levels of revenues $r_1 = r_2$, traffic intensity $(\lambda_1 + \lambda_2)$, and choice parameters to represent low, medium, high revenues, traffic intensity levels, and cross lead-time elasticities, respectively. The lead-time sensitivity of product 1 is taken constant at $d_1^{\max} = 8$ and for product 2, five levels of lead-time sensitivity are tested, $d_2^{\max} \in \{4, 6, 8, 10, 12\}$.

Observation 13 *As lead-time sensitivity decreases (i.e., $d^{\max}$ increases), the benefit of dynamic quotation and/or substitution decreases.*



(a) No-substitution (b) Substitution

Figure 24: The impact of $\lambda_1 + \lambda_2$ on benefits obtained by Dd_1 and/or Dd_2 w/ or w/o substitution on NS-Sd1Sd2.

Table 17: Numerical analysis settings in Section A.3.1

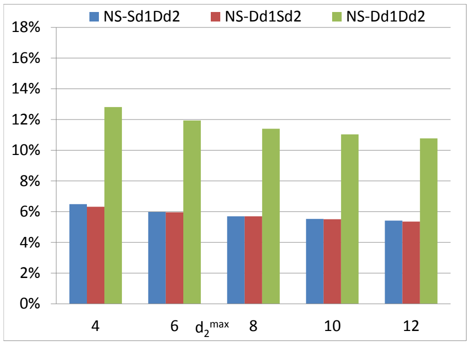
Revenues	Arrival Rates
$r_1 = r_2 \in \{5, 15, 25\}$	$(\lambda_1 + \lambda_2) \in \{0.3, 0.6, 0.9\}$ $\lambda_1/(\lambda_1 + \lambda_2) \in \{50\%\}$
Lead-time sensitivity	**Choice prob. Parameters**
$d_1^{max} = 8$ $d_2^{max} \in \{4, 6, 8, 10, 12\}$	$k_{11} = k_{22} \in \{0.15, 0.30, 0.45\}$ $k_{12} = k_{21} \in \{0.15, 0.30, 0.45\}$

Dynamic quotation and/or substitution increases the profits more under higher lead-time sensitivity (i.e., lower values of $d_2^{\max}$) as seen in Figure 25. The improvements are much higher when both products are quoted dynamically (NS-Dd1Dd2 and S-Dd1Dd2) compared to one of the products quoted static other quoted dynamic (NS-Sd1Dd2, NS-Dd1Sd2 and S-Sd1Dd2, S-Dd1Sd2).

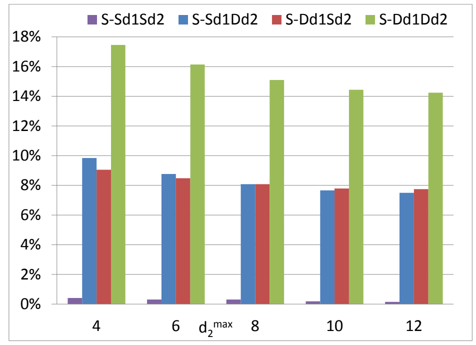
Effects of dynamic quotation, substitution and synergy of both decreases as lead-time sensitivity decreases as shown in Figure 26. The benefits obtained by substitution are insignificant since revenues and arrival rates are equal for both products which makes substitution less favorable.

A.3.2 Impact of Choice Parameters

In this section, we test the impact of cross lead-time elasticity with numerical analysis settings as given in Table 18. We test three levels of $r_1 = r_2$ and $\lambda_1 + \lambda_2$ to represent

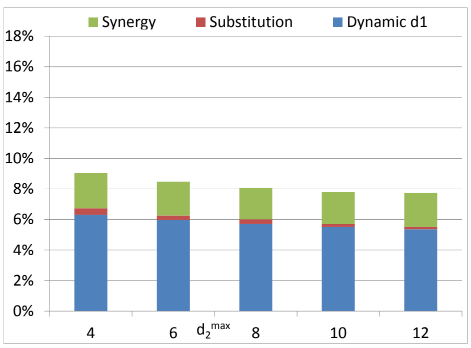


(a) No-substitution

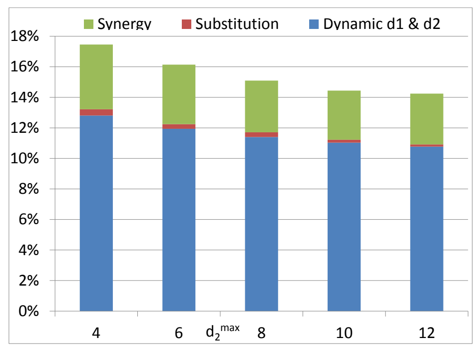


(b) Substitution

Figure 25: The impact of $d_2^{\max}$ on benefits obtained by Dd_1 and/or Dd_2 w/ or w/o substitution on NS-Sd1Sd2.



(a) Dynamic d_1 or d_2



(b) Dynamic d_1 and d_2

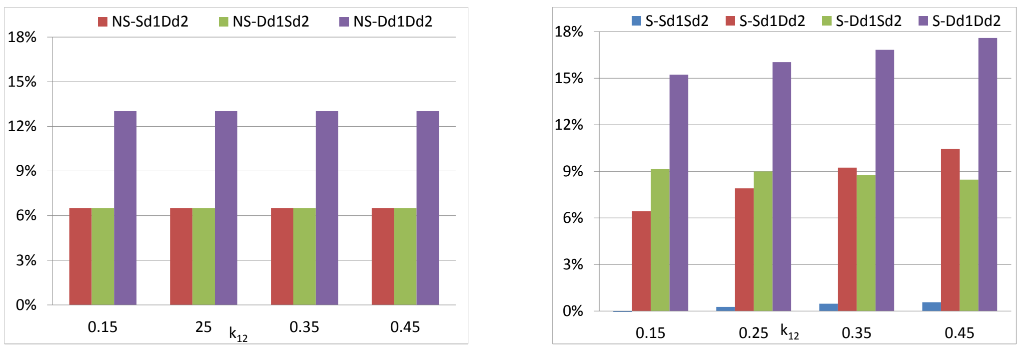
Figure 26: The impact of $d_2^{\max}$ on benefits obtained by Dd_1 (Dd_1Dd_2), substitution and synergy of both on NS-Sd1Sd2.

low, medium, high revenues and traffic intensities, two levels of $d^{\max}$ to represent low and high lead-time sensitivity for both products. k_{12} which is the cross lead-time elasticity in $p_{11}(d_1, d_2)$ (i.e., probability of customer type 1 choosing product 1) is varied from 0.15 to 0.45 while k_{11} is taken constant at 0.15 and three levels are tested for customer 2, i.e., $k_{21} = k_{22} = \in \{0.15, 0.30, 0.45\}$.

Observation 14 *The benefits obtained by dynamic quotation of product 2 and substitution as well as dynamic quotation of both products and substitution increase as cross lead-time elasticity for $p_{11}(d_1, d_2)$ (i.e., k_{12}) increases. On the other hand, the*

Table 18: Numerical analysis settings in Section A.3.2

Revenues	Arrival Rates
$r_1 = r_2 \in \{5, 15, 25\}$	$(\lambda_1 + \lambda_2) \in \{0.3, 0.6, 0.9\}$ $\lambda_1/(\lambda_1 + \lambda_2) \in \{50\%\}$
Lead-time sensitivity	**Choice prob. Parameters**
$d_1^{max} = d_2^{max} \in \{4, 8\}$	$k_{12} \in \{0.15, 0.25, 0.35, 0.45\}$ $k_{11} = 0.15$ $k_{21} = k_{22} \in \{0.15, 0.30, 0.45\}$

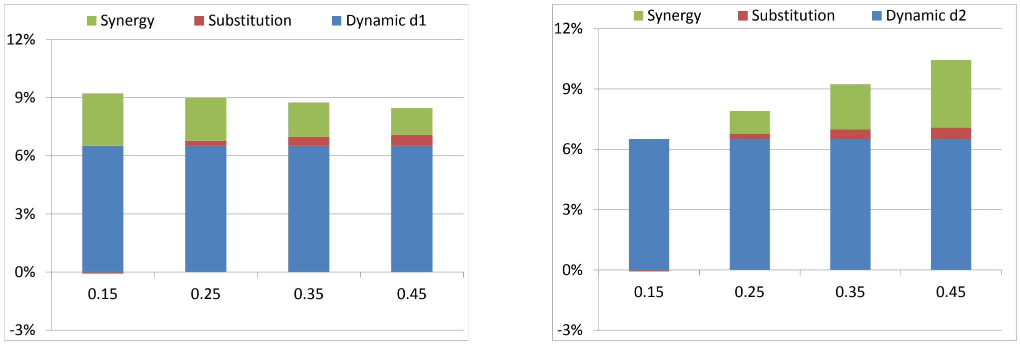


(a) No-substitution (b) Substitution

Figure 27: The impact of k_{12} on improvements obtained by dynamic lead-time quotation w/ and w/o substitution.

benefits obtained by dynamic quotation of product 1 *and substitution decreases as* k_{12} *increases.*

Figure 27 depicts the improvements on base scenario by dynamic quotation of lead-times (and substitution). Without substitution, change in k_{12} does not affect the profits since other product's lead-time does not affect acceptance probability when there is no substitution. When product 2 is quoted dynamically, S-Sd1Dd2 and S-Dd1Dd2, the benefits obtained increase as cross lead-time elasticity increases. By quoting d_2 dynamically, $p_{11}(d_1, d_2)$ is affected more as cross lead-time elasticity increases. However, if d_1 is quoted dynamically with substitution, S-Dd1Sd2, as k_{12} increases the benefits obtained decrease. In this case, since d_2 is quoted static; the benefits obtained decrease as coefficient of the static quoted lead-time increases.



(a) Dynamic d_1 (b) Dynamic d_2

Figure 28: The impact of k_{12} on benefits obtained by Dd_1 (Dd_2), substitution and synergy of both on NS-Sd1Sd2.

Figure 28 depicts the magnitudes of benefits obtained by substitution, dynamic quotation and the synergistic effects applied on base scenario, NS-Sd1Sd2. Dynamic quotation of d_1 or d_2 (NS-Dd1Sd2 or NS-Sd1Dd2) results in 6.51% improvement in profit compared to NS-Sd1Sd2, and improvement does not change with changes in k_{12}. Benefits obtained by substitution (S-Sd1Sd2) vary with k_{12}, for $k_{12} = 0.15$ the substitution results in decrease in profits. This is an unexpected result but the reason is the choice of acceptance probability function for no-substitution case as explained in subsection 2.5.1. Acceptance probability for no-substitution case assumes that the other product's lead-time is quoted at maximum value compared to choice probability in substitution case. As k_{12} increases, the synergistic effects decrease when d_1 is quoted dynamic, and increase when d_2 is quoted dynamic.

A.3.3 General Choice Model

General choice model for n products is defined as follows for any customer type for preferred products and other substitutable products where p_{ci} denotes the choice probability for preferred product i for customer type c and p_{cj} denoted the one for other substitutable products j for customer type c:

$$p_{ci}(d_1, d_2, ..., d_n) = \left[-\sum_{j=1}^{n} a_{cj}(d_j^{\max} - d_j) \right]^+, \tag{58}$$

where $a_{cj} = \frac{k_{cj}}{d_j^{\max}}$ for $j \neq i$ is the cross lead-time elasticity for choice probability of customer c for product j and $a_{ci} = -\frac{1}{d_i^{\max}}$ if i is the preferred product for customer type c, and $a_{ci} = -\frac{k_{ci}+k_{cp}}{d_i^{\max}}$ if i is any other substitutable product for customer type c.

A.4 Proofs

Proof. Proof of Lemma 1. Using Equation (6) and Leibniz integral rule,

$$\frac{\partial L(x_i, d_i)}{\partial d_i} = \int_{d_i}^{\infty} \frac{\partial (y - d_i) h_{x_i+1}(y)}{\partial d_i} dy = \int_{d_i}^{\infty} -h_{x_i+1}(y) dy \tag{59}$$

where $h_{x_i+1}(.)$ is the probability density function of Erlang distribution with parameters μ_i and $x_i + 1$. Let $H_{x_i+1}(.)$ denote the cumulative distribution of same distribution, i.e., $H_{x_i+1}(y) = \int_0^y h_{x_i+1}(y)dy$. Hence,

$$\frac{\partial L(x_i, d_i)}{\partial d_i} = H_{x_i+1}(d_i) - 1 < 0 \textit{ since } H_{x_i+1}(.) < 1 \ \forall \ d \geq 0. \tag{60}$$

Therefore, $L(x_i, d_i)$ is decreasing in $d_i \geq 0$.

The second derivative is non-negative (due to the fact that $h_{x_i+1}(.)$ is a probability density function and non-negative) which proves convexity of $L(x_i, d_i)$:

$$\frac{\partial^2 L(x_i, d_i)}{\partial d_i^2} = h_{x_i+1}(d_i) \geq 0 \textit{ since } h_{x_i+1}(.) \geq 0 \ \forall \ d_i \geq 0. \tag{61}$$

■

Proof. Proof of Proposition 1. $\alpha_{x_1,x_2}(d_1)$ and $\beta_{x_1,x_2}(d_2)$ are increasing and concave in d_1 and d_2 respectively by Lemma 1.

Assume $(d_1^*, d_2^*) \notin \Gamma_{x_1,x_2}$ and $(d_1^*, d_2^*) \neq (d_1^{\max}, d_2^{\max})$ by contradiction. Then, $\alpha_{x_1,x_2}(d_1^*, d_2^*) < 0$ and $\beta_{x_1,x_2}(d_1^*, d_2^*) < 0$ which results in $\Pi^*_{x_1,x_2}(d_1, d_2) < 0$ since $p_{ci}(d_1, d_2) > 0$ for

$d_1 < d_1^{\max}$ and $d_2 < d_2^{\max}$. However, a natural lower bound on $\Pi^*_{x_1,x_2}(d_1,d_2)$ is 0 which occurs at $(d_1^*, d_2^*) = (d_1^{\max}, d_2^{\max})$. ■

Proof. Proof of Lemma 2. We will describe structural properties of relative value function in the infinite horizon problem, $v^*(x_1,x_2)$, in terms of state variables x_1 and x_2. We first define some additional notation, arrival and departure operators, to shorten the analysis. We will use the finite horizon version of the same problem, where $J_n(x_1,x_2)$ denote the total profit of the system starting in state (x_1,x_2) with n transitions remaining in the future and the optimality equation of the finite horizon problem is given as:

$$J_{n+1}(x_1,x_2) = T_{ARR}J_n(x_1,x_2) + \frac{\mu_1}{\nu}T_{DEP1}J_n(x_1,x_2) + \frac{\mu_2}{\nu}T_{DEP2}J_n(x_1,x_2), \quad (62)$$

where,

$$\begin{aligned}
T_{DEP1}J(x_1,x_2) &= J((x_1-1)^+, x_2) && (63)\\
T_{DEP2}J(x_1,x_2) &= J(x_1, (x_2-1)^+) && (64)\\
T_{ARR}J(x_1,x_2) &= \max_{d_1,d_2}\Bigg\{\left(\frac{\lambda_1}{\nu}p_{10}(d_1,d_2) + \frac{\lambda_2}{\nu}p_{20}(d_1,d_2)\right)J(x_1,x_2)\\
&+ \left(\frac{\lambda_1}{\nu}p_{12}(d_1,d_2) + \frac{\lambda_2}{\nu}p_{22}(d_1,d_2)\right)(r_2 - l_2L(x_2,d_2) + J(x_1,x_2+1))\\
&+ \left(\frac{\lambda_1}{\nu}p_{11}(d_1,d_2) + \frac{\lambda_2}{\nu}p_{21}(d_1,d_2)\right)(r_1 - l_1L(x_1,d_1) + J(x_1+1,x_2))\Bigg\} && (65)
\end{aligned}$$

We will prove that the operators, T_{DEP1}, T_{DEP2} and T_{ARR}, preserve the monotonicity properties of the function on which they are applied, $J(x_1,x_2)$. We prove that following monotonicity equations hold for these operators:

$$\textit{Monotonicity in } x_1\textit{:} \quad TJ(x_1,x_2) \geq TJ(x_1+1,x_2) \quad (66)$$

$$\textit{Monotonicity in } x_2\textit{:} \quad TJ(x_1,x_2) \geq TJ(x_1,x_2+1) \quad (67)$$

We will show the monotonicity of operators in x_1 and one can show similarly that operators are monotone in x_2, therefore skipped.

- **Monotonicity of T_{DEP1}:**

  Assuming $J(x_1, x_2)$ is monotone in x_1, we will show that monotonicity is preserved under operator T_{DEP1}, i.e., satisfying $T_{DEP1}J(x_1, x_2) \geq T_{DEP1}J(x_1 + 1, x_2)$.

$$T_{DEP1}J(x_1,x_2) \geq T_{DEP1}J(x_1+1,x_2) = \begin{cases} J(x_1-1,x_2) \geq J(x_1,x_2), & \text{for } x_1 > 0 \quad x_2 \geq 0 \\ J(0,x_2) \geq J(0,x_2), & \text{for } x_1 = 0 \quad x_2 \geq 0. \end{cases} \tag{68}$$

  First case is true by monotonicity of $J(x_1, x_2)$ in x_1, and the right hand side and the left hand side are equal in the second case. Therefore, we can conclude that T_{DEP1} operator preserves the monotonicity of $J(x_1, x_2)$ in x_1.

- **Monotonicity of T_{DEP2}:**

  Assuming $J(x_1, x_2)$ is monotone in x_1, we will show that monotonicity is preserved under operator T_{DEP2}, i.e., satisfying $T_{DEP2}J(x_1, x_2) \geq T_{DEP2}J(x_1 + 1, x_2)$.

$$T_{DEP2}J(x_1,x_2) \geq T_{DEP2}J(x_1+1,x_2) = \begin{cases} J(x_1,x_2-1) \geq J(x_1+1,x_2-1), & \text{for } x_1 \geq 0 \quad x_2 > 0 \\ J(x_1,0) \geq J(x_1+1,0), & \text{for } x_1 \geq 0 \quad x_2 = 0. \end{cases} \tag{69}$$

  Both cases are true by monotonicity of $J(x_1, x_2)$ in x_1, and we can conclude that T_{DEP2} operator preserves the monotonicity of $J(x_1, x_2)$ in x_1.

- **Monotonicity of T_{ARR}:**

  Assuming $J(x_1, x_2)$ is monotone in x_1, we will show that monotonicity is preserved under operator T_{ARR}, i.e., satisfying $T_{ARR}J(x_1, x_2) \geq T_{ARR}J(x_1+1, x_2)$. Let (d_1^{00}, d_2^{00}) and (d_1^{10}, d_2^{10}) be the optimal lead-time quotes at states (x_1, x_2) and $(x_1 + 1, x_2)$, respectively. Then, since (d_1^{00}, d_2^{00}) are the optimal lead-time quotes at state (x_1, x_2), we can write the following inequality for state (x_1, x_2).

$$\begin{aligned}
&\left\{\left(\frac{\lambda_1}{\nu}p_{11}(d_1^{00}, d_2^{00}) + \frac{\lambda_2}{\nu}p_{21}(d_1^{00}, d_2^{00})\right)(r_1 - l_1L(x_1, d_1^{00}) + J(x_1+1, x_2))\right.\\
&\quad + \frac{\lambda_1}{\nu}p_{12}(d_1^{00}, d_2^{00}) + \frac{\lambda_2}{\nu}p_{22}(d_1^{00}, d_2^{00})\Big)(r_2 - l_2L(x_2, d_2^{00}) + J(x_1, x_2+1))\\
&\qquad \left. + \left(\frac{\lambda_1}{\nu}p_{10}(d_1^{00}, d_2^{00}) + \frac{\lambda_2}{\nu}p_{20}(d_1^{00}, d_2^{00})\right)J(x_1, x_2)\right\}\\
&\geq\\
&\left\{\left(\frac{\lambda_1}{\nu}p_{11}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{21}(d_1^{10}, d_2^{10})\right)(r_1 - l_1L(x_1, d_1^{10}) + J(x_1+1, x_2))\right.\\
&\quad + \frac{\lambda_1}{\nu}p_{12}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{22}(d_1^{10}, d_2^{10})\Big)(r_2 - l_2L(x_2, d_2^{10}) + J(x_1, x_2+1))\\
&\qquad \left. + \left(\frac{\lambda_1}{\nu}p_{10}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{20}(d_1^{10}, d_2^{10})\right)J(x_1, x_2)\right\}. \qquad (70)
\end{aligned}$$

Also, by monotonicity of $J(x_1, x_2)$ in x_1 and taking into account that $L(x_1 + 1, d_1) = L(x_1, d_1) + \frac{e^{-\mu_1 d_1}}{\mu_1}\sum_{k=0}^{x_1+1}\frac{(d_1\mu_1)^k}{k!}$, (i.e., $L(x_1+1, d_1) > L(x_1, d_1)$); the right hand side of Equation (70) is less than $T_{ARR}J(x_1+1, x_2)$.

$$\begin{aligned}
&\left\{\left(\frac{\lambda_1}{\nu}p_{11}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{21}(d_1^{10}, d_2^{10})\right)(r_1 - l_1L(x_1, d_1^{10}) + J(x_1+1, x_2))\right.\\
&\quad + \frac{\lambda_1}{\nu}p_{12}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{22}(d_1^{10}, d_2^{10})\Big)(r_2 - l_2L(x_2, d_2^{10}) + J(x_1, x_2+1))\\
&\qquad \left. + \left(\frac{\lambda_1}{\nu}p_{10}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{20}(d_1^{10}, d_2^{10})\right)J(x_1, x_2)\right\}\\
&\geq\\
&\left\{\left(\frac{\lambda_1}{\nu}p_{11}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{21}(d_1^{10}, d_2^{10})\right)(r_1 - l_1L(x_1+1, d_1^{10}) + J(x_1+2, x_2))\right.\\
&\quad + \frac{\lambda_1}{\nu}p_{12}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{22}(d_1^{10}, d_2^{10})\Big)(r_2 - l_2L(x_2, d_2^{10}) + J(x_1+1, x_2+1))\\
&\qquad \left. + \left(\frac{\lambda_1}{\nu}p_{10}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{20}(d_1^{10}, d_2^{10})\right)J(x_1+1, x_2)\right\}. \qquad (71)
\end{aligned}$$

Now by combining Equations (70) and (71), we obtain Equation (72). Hence, we conclude that monotonicity inequality, $T_{ARR}J(x_1, x_2) \geq T_{ARR}J(x_1+1, x_2)$, holds and the operator T_{ARR} preserves the monotonicity of $J(x_1, x_2)$ in x_1.

$$\begin{aligned}
&\Bigg\{\Bigg(\frac{\lambda_1}{\nu}p_{11}(d_1^{00}, d_2^{00}) + \frac{\lambda_2}{\nu}p_{21}(d_1^{00}, d_2^{00})\Bigg)(r_1 - l_1L(x_1, d_1^{00}) + J(x_1+1, x_2)) \\
&\quad + \frac{\lambda_1}{\nu}p_{12}(d_1^{00}, d_2^{00}) + \frac{\lambda_2}{\nu}p_{22}(d_1^{00}, d_2^{00})\Bigg)(r_2 - l_2L(x_2, d_2^{00}) + J(x_1, x_2+1)) \\
&\quad + \Bigg(\frac{\lambda_1}{\nu}p_{10}(d_1^{00}, d_2^{00}) + \frac{\lambda_2}{\nu}p_{20}(d_1^{00}, d_2^{00})\Bigg)J(x_1, x_2)\Bigg\} \\
&\geq \\
&\Bigg\{\Bigg(\frac{\lambda_1}{\nu}p_{11}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{21}(d_1^{10}, d_2^{10})\Bigg)(r_1 - l_1L(x_1+1, d_1^{10}) + J(x_1+2, x_2)) \\
&\quad + \frac{\lambda_1}{\nu}p_{12}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{22}(d_1^{10}, d_2^{10})\Bigg)(r_2 - l_2L(x_2, d_2^{10}) + J(x_1+1, x_2+1)) \\
&\quad + \Bigg(\frac{\lambda_1}{\nu}p_{10}(d_1^{10}, d_2^{10}) + \frac{\lambda_2}{\nu}p_{20}(d_1^{10}, d_2^{10})\Bigg)J(x_1+1, x_2)\Bigg\}. \quad (72)
\end{aligned}$$

We show that all operators preserve the monotonicity properties of $J(x_1, x_2)$ in x_1 and x_2; hence $J_{n+1}(x_1, x_2)$ is monotone in x_1 and x_2 given that it is monotone in x_1 and x_2. By taking the limit as $n \to \infty$, we obtain $\lim_{n\to\infty} J_n(x_1, x_2) = J^*(x_1, x_2)$ where $J^*(x_1, x_2)$ is the total profit of the system starting in state (x_1, x_2) in the infinite horizon version of the problem. Therefore structural results obtained by $J_n(x_1, x_2)$ hold for $J^*(x_1, x_2)$. Note that, Equation (62) can be viewed as the DP algorithm for a finite horizon problem with terminal profit function equal to $J_0(x_1, x_2)$ by reversing the time index; and $J_n(x_1, x_2)$ is the optimal profit starting from state (x_1, x_2) of a n-stage problem with $J_0(x_1, x_2)$ as terminal profit at the end of n-stages (Bertsekas [4]). Moreover, these structural results are also valid for the average reward criteria version of the problem with relative value functions $v^*(x_1, x_2)$. ■

Proof. Proof of Lemma 3. First analyze the behavior of $\Pi_{x_1,x_2}(d_1, d_2)$ w.r.t. d_1. Assuming $b_{11} > b_{21}$ i.e., $k_{12}k_{22} + k_{12}k_{21} - k_{22} < 0$, the regions will be defined as follows and analysis of each region is given below:

1. Region A: $0 < d_1 \leq b_{21}$

$$\begin{aligned} \frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial {d_1}^2} &= -2\frac{\lambda_1+\lambda_2(k_{21}+k_{22})}{\nu d_1^{\max}}(-l_1\frac{\partial L(x_1,d_1)}{\partial d_1}) \\ &+ \frac{\lambda_1 p_{11}(d_1,d_2)+\lambda_2 p_{21}(d_1,d_2)}{\nu}(-l_1\frac{\partial^2 L(x_1,d_1)}{\partial {d_1}^2}) \end{aligned} \quad (73)$$

$\frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial {d_1}^2} \leq 0$ since $L(x_i,d_i)$ is convex and decreasing in $d_i > 0$. $\Pi_{x_1,x_2}(d_1,d_2)$ is concave in d_1 and the maximum occurs either at boundaries or the point where first derivative equals 0.

$$\begin{aligned} \frac{\partial \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_1} = &-\frac{\lambda_1+\lambda_2(k_{21}+k_{22})}{\nu d_1^{\max}}\alpha_{x_1,x_2}(d_1) + \frac{\lambda_1 k_{11}+\lambda_2 k_{21}}{\nu d_1^{\max}}\beta_{x_1,x_2}(d_2) \\ &+\frac{\lambda_1 p_{11}(d_1,d_2)+\lambda_2 p_{21}(d_1,d_2)}{\nu}(-l_1\frac{\partial L(x_1,d_1)}{\partial d_1}) = 0 \end{aligned} \quad (74)$$

2. Region B : $b_{21} \leq d_1 \leq b_{11}$, i.e., $p_{21}(d_1,d_2) = 0$

Again $\frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial {d_1}^2} \leq 0$, and $\Pi_{x_1,x_2}(d_1,d_2)$ is concave in d_1 and the maximum occurs either at boundaries or the point where first derivative equals 0.

$$\begin{aligned} \frac{\partial \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_1} = &-\frac{\lambda_1}{\nu d_1^{\max}}\alpha_{x_1,x_2}(d_1) + \frac{\lambda_1 p_{11}(d_1,d_2)}{\nu}(-l_1\frac{\partial L(x_1,d_1)}{\partial d_1}) \\ &+\frac{\lambda_1 k_{11}+\lambda_2 k_{21}}{\nu d_1^{\max}}\beta_{x_1,x_2}(d_2) = 0 \end{aligned} \quad (75)$$

3. Region C : $b_{11} \leq d_1 \leq d_1^{\max}$, i.e., $p_{21}(d_1,d_2) = p_{11}(d_1,d_2) = 0$

$\frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial {d_1}^2} = 0$, and $\Pi^*_{x_1,x_2}(d_1,d_2)$ is linear in d_1 and sign of $\beta_{x_1,x_2}(d_2)$ determines whether $\Pi^*_{x_1,x_2}(d_1,d_2)$ is increasing or decreasing in d_1

$$\frac{\partial \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_1} = \frac{\lambda_1 k_{11}+\lambda_2 k_{21}}{\nu d_1^{\max}}\beta_{x_1,x_2}(d_2) \quad (76)$$

The analysis of behavior of $\Pi_{x_1,x_2}(d_1,d_2)$ w.r.t. d_2 is performed for three regions by assuming $b_{22} > b_{12}$ i.e., $k_{12}k_{21} + k_{11}k_{21} - k_{11} < 0$.

1. Region I : $0 < d_2 \leq b_{12}$:

$$\begin{aligned} \frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial {d_2}^2} &= -2\frac{\lambda_1(k_{11}+k_{12})+\lambda_2}{\nu d_2^{\max}}(-l_2\frac{\partial L(x_2,d_2)}{\partial d_2}) \\ &+ \frac{\lambda_1 p_{12}(d_1,d_2)+\lambda_2 p_{22}(d_1,d_2)}{\nu}(-l_2\frac{\partial^2 L(x_2,d_2)}{\partial {d_2}^2}) \end{aligned} \quad (77)$$

$\frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_2{}^2} \leq 0$ since $L(x_i, d_i)$ is convex and decreasing in $d_i > 0$. $\Pi_{x_1,x_2}(d_1, d_2)$ is concave in d_2 and the maximum occurs either at boundaries or the point where first derivative equals 0.

$$\frac{\partial \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_2} = \frac{\lambda_1 k_{12} + \lambda_2 k_{22}}{\nu d_2^{\max}} \alpha_{x_1,x_2}(d_1) - \frac{\lambda_1(k_{11}+k_{12}) + \lambda_2}{\nu d_2^{\max}} \beta_{x_1,x_2}(d_2) + \frac{\lambda_1 p_{12}(d_1,d_2) + \lambda_2 p_{22}(d_1,d_2)}{\nu}(-l_2 \frac{\partial L(x_2,d_2)}{\partial d_2}) = 0 \quad (78)$$

2. Region B :$b_{12} \leq d_2 \leq b_{22}$, i.e., $p_{12}(d_1, d_2) = 0$

   Again, $\frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_2{}^2} \leq 0$, and $\Pi^*_{x_1,x_2}(d_1, d_2)$ is concave in d_2 and the maximum occurs either at boundaries or the point where first derivative equals 0.

$$\frac{\partial \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_2} = \frac{\lambda_1 k_{12} + \lambda_2 k_{22}}{\nu d_2^{\max}} \alpha_{x_1,x_2}(d_1) - \frac{\lambda_2}{\nu d_2^{\max}} \beta_{x_1,x_2}(d_2) + \frac{\lambda_2 p_{22}(d_1,d_2)}{\nu}(-l_2 \frac{\partial L(x_2,d_2)}{\partial d_2}) = 0 \quad (79)$$

3. Region C : $b_{22} \leq d_2 \leq d_2^{\max}$, i.e., $p_{12}(d_1, d_2) = p_{22}(d_1, d_2) = 0$

   $\frac{\partial^2 \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_2{}^2} = 0$, and $\Pi^*_{x_1,x_2}(d_1, d_2)$ is linear in d_2 and sign of $\alpha_{x_1,x_2}(d_1)$ determines whether $\Pi^*_{x_1,x_2}(d_1, d_2)$ is increasing or decreasing in d_2

$$\frac{\partial \Pi_{x_1,x_2}(d_1,d_2)}{\partial d_2} = \frac{\lambda_1 k_{12} + \lambda_2 k_{22}}{\nu d_2^{\max}} \alpha_{x_1,x_2}(d_1) \quad (80)$$

■

Proof. Proof of Theorem 1. Proof of Theorem 1 requires the following Lemma which states that the optimal lead-time quote can not be the boundary point of RegionS B and C; further if optimal is in Region C, then it is the maximum lead-time quote. For a given d_2 value, if the optimum d_1 is in the region that the choice probability of product 1 is zero, then it must be $d_1^{\max}$. Since the expected profit from product 1 will be zero, increasing d_1 will only increase the expected profit from product 2 and the overall profit.

Lemma 8 *If $d_1^* \geq b_{21}(d_2^*)$, then $d_1^* = d_1^{\max}$. Similarly, if $d_2^* \geq b_{22}(d_1^*)$, then $d_2^* = d_2^{\max}$.*

Proof. Proof of Lemma 8. If $d_1^* \geq b_{21}(d_2^*)$, then optimum is either $b_{21}(d_2^*)$ or $d_1^{\max}$ since $\Pi^*_{x_1,x_2}(d_1, d_2)$ is linear in Region C w.r.t. d_1 and further the sign of $\beta_{x_1,x_2}(d_2^*)$ determines whether $\Pi^*_{x_1,x_2}(d_1, d_2)$ is increasing or decreasing by Lemma 3.

- If $\beta_{x_1,x_2}(d_2^*) > 0$, $\Pi^*_{x_1,x_2}(d_1, d_2)$ is increasing for $d_1^* > b_{21}(d_2^*)$ and optimum is $d_1^{\max}$.

- If $\beta_{x_1,x_2}(d_2^*) < 0$ and $\alpha_{x_1,x_2}(d_1^*) < 0$, optimum is $(d_1^*, d_2^*) = (d_1^{\max}, d_2^{\max})$ by Proposition 1.

- If $\beta_{x_1,x_2}(d_2^*) < 0$ and $\alpha_{x_1,x_2}(d_1^*) > 0$, then $\frac{\partial \Pi^*_{x_1,x_2}(d_1,d_2)}{\partial d_2} > 0$ in all regions according to the proof of Lemma 3 which contradicts with d_2^* being optimal.

The above results are sufficient and if $d_1^* \geq b_{21}(d_2^*)$, then $d_1^* = d_1^{\max}$. The proof for d_2^* is similar, and hence omitted. ■

$\alpha_{x_1,x_2}(d_1)$ and $\beta_{x_1,x_2}(d_2)$ are increasing and concave in d_1 and d_2 respectively by Lemma 1. Then, $\alpha_{x_1,x_2}(d_1^{\max}) < 0$ implies that $\alpha_{x_1,x_2}(d_1) < 0$ $\forall d_1 \in (0, d_1^{\max}]$ and similarly $\beta_{x_1,x_2}(d_2^{\max}) < 0$ implies that $\beta_{x_1,x_2}(d_2) < 0$ $\forall d_2 \in (0, d_2^{\max}]$.

(a) If $\alpha_{x_1,x_2}(d_1^{\max}) < 0$, then $\alpha_{x_1,x_2}(d_1) < 0$ $\forall d_1 \in (0, d_1^{\max}]$ since $\alpha_{x_1,x_2}(d_1)$ is increasing and concave in d_1 by Lemma 1. Given $\alpha_{x_1,x_2}(d_1) < 0$, $\frac{\partial \Pi^*_{x_1,x_2}(d_1,d_2)}{\partial d_1} > 0$ in all regions according to the proof of Lemma 3. Therefore, the optimal is not in Regions A and B (the equalities (74) and (75) do not hold) and is in Region C and is $d_1^{\max}$ since $\Pi^*_{x_1,x_2}(d_1, d_2)$ is increasing in Region C.

(b) Else, i.e., $\alpha_{x_1,x_2}(d_1^{\max}) > 0$:

(i) if $\beta_{x_1,x_2}(d_2^{\max}) < 0$, $\beta_{x_1,x_2}(d_2) < 0$ $\forall d_2 \in (0, d_2^{\max}]$ since $\beta_{x_1,x_2}(d_2)$ is increasing and concave in d_2 by Lemma 1. Since $\alpha_{x_1,x_2}(d_1^{\max}) > 0$, $\alpha_{x_1,x_2}(d_1^*) > 0$ must hold.

$\alpha_{x_1,x_2}(d_1^*) > 0$ and $\beta_{x_1,x_2}(d_2) < 0$ $\forall d_2$ implies that $\frac{\partial \Pi^*_{x_1,x_2}(d_1,d_2)}{\partial d_2} > 0$, hence $d_2^* = d_2^{\max}$. For $d_2^* = d_2^{\max}$, only Region A exists and is concave for $d_1 \in (0, d_1^{\max}]$. The optimum is either ϵ or the maximizer of $\frac{\partial \Pi^*_{x_1,x_2}(d_1,d_2)}{\partial d_1}$.

- if $\frac{\partial \Pi^*}{\partial d_1}|_{d_1=\epsilon} < 0$, then $d_1^* = \epsilon$.
- Otherwise, $\exists d_1^*$ s.t. $\frac{\partial \Pi^*}{\partial d_1}|_{(d_1^*,d_2^{\max})} = 0$.

(ii) else, i.e., $\beta_{x_1,x_2}(d_2^{\max}) > 0$, both $\alpha_{x_1,x_2}(d_1)$ and $\beta_{x_1,x_2}(d_2)$ can be either positive or negative in this case. Therefore, the optimum can be any of the possible values, ϵ, maximum of Region A or B, boundary of regions A and B and $d_1^{\max}$. By Lemma 8, the boundary of regions B and C cannot be optimal; and the proof is complete.

■

APPENDIX B

PROOFS IN CHAPTER 3

Proof of Lemma 5. A closed-form expression is derived for $L(i,d)$ using the methodology of Hafizoglu et al. [26].

$$L(i,d) = \int_d^\infty (t-d) g_i(t) dt, \tag{81}$$

where $g_i(\cdot)$ denotes the probability density function (pdf) of time in system (TIS) of an incoming order when there are i orders in the system.

Consider the marked order observing i orders in front upon arrival, where $i > 0$. Let $G_i(\cdot)$ denote the cumulative distribution for TIS of an incoming order when there are i orders in the system assuming that the incoming order is not cancelled. One can rewrite $G_i(\cdot)$ by conditioning on the next event, considering two possibilities:

1. Cancelation of an order in front with rate $i\gamma$, pulling the marked order one position front,

2. Service completion of the order currently in service with rate μ, pulling the marked order one position to the front,

Hence, by conditioning on the time and type of the next event we reach,

$$\begin{aligned} G_i(t) &= \frac{i\gamma}{\mu + i\gamma} \int_0^t G_{i-1}(t-x)(\mu + i\gamma) e^{-(\mu+i\gamma)x} dx \\ &+ \frac{\mu}{\mu + i\gamma} \int_0^t G_{i-1}(t-x)(\mu + i\gamma) e^{-(\mu+i\gamma)x} dx \\ &= \int_0^t G_{i-1}(t-x)(\mu + i\gamma) e^{-(\mu+i\gamma)x} dx. \end{aligned} \tag{82}$$

Note that, the integrals in the first and second terms are convolutions of the exponential distribution, and the conditional cumulative TIS distributions.

Let $\widetilde{g}_i(s)$ denote the Laplace transform of $g_i(\cdot)$. We need to derive $\widetilde{g}_i(s)$ using Laplace-Stieltjes Transforms (LST) of the cumulative distribution functions, $G_i(\cdot)$. The LST of $G_i(\cdot)$ is equivalent to the sought Laplace Transform, $g_i(\cdot)$ by the following LST definition (Heyman and Sobel [27]):

$$\int_0^\infty e^{-sx} dG_i(x)dx = \int_0^\infty e^{-sx} g_i(x)dx = \widetilde{g}_i(s), \quad s \in \mathbb{C}. \tag{83}$$

Taking the LST's of both sides in Equation (82), and using the convolution property, we get

$$\widetilde{g}_i(s) = \widetilde{g}_{i-1}(s) \frac{\mu + i\gamma}{s + \mu + i\gamma} \quad s \in \mathbb{C}. \tag{84}$$

Using the convolution property, we obtain

$$g_i(t) = \int_0^t g_{i-1}(t-x)(\mu + i\gamma)e^{-(\mu+i\gamma)x} dx. \tag{85}$$

For $i = 0$, we have

$$g_0(t) = \mu e^{-\mu t}. \tag{86}$$

By arranging the terms, we reach the recursive form:

$$g_i(t) = \frac{e^{-\mu t}}{\gamma^i i!}(1 - e^{-\gamma t})^i \prod_{j=0}^{i} (\mu + j\gamma). \tag{87}$$

Then using the definition of $L_i(d)$ and plugging in closed-form expression of $g_i(t)$, $L_i(d)$ becomes as follows:

$$L_i(d) = \frac{1}{\gamma^i i!} \prod_{j=0}^{i} (\mu + j\gamma) \int_d^\infty (t-d)e^{-\mu t}(1 - e^{-\gamma t})^i dt. \tag{88}$$

Then by recursion, closed-form expression for $L_i(d)$ is derived:

$$L_i(d) = \frac{1}{\gamma^i i!} \prod_{j=0}^{i} (\mu + j\gamma) \sum_{k=0}^{i} (-1)^k \binom{i}{k} \frac{e^{-(\mu+k\gamma)d}}{(\mu + k\gamma)^2}. \tag{89}$$

$\square$

Proof of Lemma 4. Let $X(i)$ be TIS of an incoming order when there are i orders in front upon arrival assuming that the incoming order is not cancelled, the pdf of $X(i)$ is $g_i(\cdot)$ for which a closed-form expression is derived in Equation (87). Let Y be time to cancellation which is exponentially distributed with rate γ. Then $\overline{q}(i)$ is the probability that order will not be cancelled, i.e., cancellation time is greater than TIS:

$$1 - q(i) = \overline{q}(i) = Pr(Y > X(i)), \tag{90}$$

where Y is exponential with rate γ and $X(i)$ has pdf $g_i(\cdot)$.

$$\begin{aligned}
\overline{q}(i) &= Pr(Y > X(i)) \\
&= \int_0^\infty P(Y > X(i) \mid X(i) = t) g_i(t) dt \\
&= \int_0^\infty P(Y > t) \frac{e^{-\mu t}}{\gamma^i i!}(1 - e^{-\gamma t})^i \prod_{j=0}^{i} (\mu + j\gamma) dt \\
&= \int_0^\infty e^{-\gamma t} \frac{e^{-\mu t}}{\gamma^i i!}(1 - e^{-\gamma t})^i \prod_{j=0}^{i} (\mu + j\gamma) dt \\
&= \frac{1}{\gamma^i i!} \prod_{j=0}^{i} (\mu + j\gamma) \int_0^\infty e^{-(\mu+\gamma)t}(1 - e^{-\gamma t})^i dt. \qquad (91)
\end{aligned}$$

Then by recursion, $\overline{q}(i)$ simplifies to:

$$\overline{q}(i) = \frac{\mu}{\mu + (i+1)\gamma}, \tag{92}$$

and probability that an incoming order will be cancelled becomes:

$$q(i) = \frac{(i+1)\gamma}{\mu + (i+1)\gamma}. \tag{93}$$

□

Proof of Lemma 6. Using Equation (88) and Leibniz integral rule,

$$\frac{\partial L(i,d)}{\partial d} = -\frac{1}{\gamma^i i!} \prod_{j=0}^{i} (\mu + j\gamma) \int_d^\infty e^{-\mu t}(1 - e^{-\gamma t})^i dt. \tag{94}$$

Then, a closed-form expression for the first derivative of $L(i,d)$ with respect to d is derived by recursion and arranging terms as follows:

$$\frac{\partial L(i,d)}{\partial d} = -\frac{e^{-\mu d}}{\gamma^i i!}\Big\{\gamma^i i! + \sum_{k=1}^{i}\prod_{t=1}^{k}(\mu+(t-1)\gamma)(1-e^{-\gamma d})^k\gamma^{i-k}\frac{i!}{k!}\Big\}. \tag{95}$$

The first derivative of $L(i,d)$ is negative for all $d \geq 0$ since $e^{-\gamma d} \leq 1$. Therefore, $L(i,d)$ is decreasing in $d \geq 0$.

The second derivative is non-negative which proves the convexity of $L(i,d)$:

$$\frac{\partial^2 L(i,d)}{\partial d^2} = \frac{\mu e^{-\mu d}}{\gamma^i i!}(1-e^{-\gamma d})^i\prod_{k=1}^{i}(\mu+k\gamma) \geq 0 \quad \forall d \geq 0. \tag{96}$$

Lastly, we prove that as system state i increases, expected tardiness duration increases. Let $\triangle L(i,d) = L(i+1,d) - L(i,d)$ be the difference in expected tardiness duration for states $(i+1)$ and i.

$$\begin{aligned}
\triangle L(i,d) &= L(i+1,d) - L(i,d) \\
&= \frac{1}{\gamma^{i+1}(i+1)!}\prod_{j=0}^{i+1}(\mu+j\gamma)\sum_{k=0}^{i+1}(-1)^k\binom{i+1}{k}\frac{e^{-(\mu+k\gamma)d}}{(\mu+k\gamma)^2} \\
&- \frac{1}{\gamma^i i!}\prod_{j=0}^{i}(\mu+j\gamma)\sum_{k=0}^{i}(-1)^k\binom{i}{k}\frac{e^{-(\mu+k\gamma)d}}{(\mu+k\gamma)^2} \\
&= \frac{1}{\gamma^i i!}\prod_{j=0}^{i}(\mu+j\gamma)\Big\{\frac{\mu+(i+1)\gamma}{\gamma(i+1)}\sum_{k=0}^{i}(-1)^k\binom{i+1}{k}\frac{e^{-(\mu+k\gamma)d}}{(\mu+k\gamma)^2} \\
&- \sum_{k=0}^{i}(-1)^k\binom{i}{k}\frac{e^{-(\mu+k\gamma)d}}{(\mu+k\gamma)^2} + \frac{\mu+(i+1)\gamma}{\gamma(i+1)}(-1)^{i+1}\frac{e^{-(\mu_(i+1)\gamma)d}}{(\mu+(i+1)\gamma)^2}\Big\} \\
&= \frac{1}{\gamma^i i!}\prod_{j=0}^{i}(\mu+j\gamma)\Big\{\sum_{k=0}^{i}(-1)^k\binom{i}{k}\frac{e^{-(\mu+k\gamma)d}}{(\mu+k\gamma)^2}\Big(\frac{i+1}{i+1-k}\frac{\mu+(i+1)\gamma}{\gamma(i+1)} - 1\Big) \\
&+ \frac{\mu+(i+1)\gamma}{\gamma(i+1)}(-1)^{i+1}\frac{e^{-(\mu_(i+1)\gamma)d}}{(\mu+(i+1)\gamma)^2}\Big\} \\
&= \frac{1}{\gamma^i i!}\prod_{j=0}^{i}(\mu+j\gamma)\Big\{\sum_{k=0}^{i}(-1)^k\binom{i}{k}\frac{e^{-(\mu+k\gamma)d}}{(\mu+k\gamma)^2}\Big(\frac{\mu+k\gamma}{(i+1-k)\gamma}\Big) \\
&+ \frac{\mu+(i+1)\gamma}{\gamma(i+1)}(-1)^{i+1}\frac{e^{-(\mu_(i+1)\gamma)d}}{(\mu+(i+1)\gamma)^2}\Big\} \\
&= \frac{1}{\gamma^{i+1}(i+1)!}\prod_{j=0}^{i}(\mu+j\gamma)\sum_{k=0}^{i+1}(-1)^k\binom{i+1}{k}\frac{e^{-(\mu+k\gamma)d}}{(\mu+k\gamma)}.
\end{aligned} \tag{97}$$

By rearranging terms and getting rid of terms with $(-1)^k$, $\triangle L(i,d)$ reduces to the following expression which is positive since $(1-e^{-\gamma d}) \geq 0$.

$$\begin{aligned} \triangle L(i,d) &= \frac{e^{-\mu d}}{\gamma^{i+1}(i+1)!(\mu+(i+1)\gamma)}\Bigg\{(i+1)!\gamma^{i+1} \\ &+ \sum_{k=0}^{i}\left(\prod_{t=0}^{k}(\mu+t\gamma)\right)(1-e^{-\gamma d})^{k+1}\gamma^{i-k}\frac{(i+1)!}{(k+1)!}\Bigg\} > 0. \end{aligned} \tag{98}$$

$\square$

Proof of Lemma 7. Similar to the proof in Chapter 2, we use the finite horizon version of the same problem and define operators to shorten the analysis. Let $h_n(i)$ be the total profit of the system starting in state i with n transitions remaining in the future and the optimality equation of the finite horizon problem is given as:

$$h_{n+1}(i) = \frac{\lambda}{\nu}T_{ARR}h_n(i) + \frac{\mu}{\nu}T_{DEP}h_n(i) + \frac{i\gamma}{\nu}T_{CAN}h_n(i) + \frac{(N-i)\gamma}{\nu}T_{FIC}h_n(i), \tag{99}$$

where,

$$\begin{aligned} T_{ARR}h(i) &= \max_d\Bigg\{f(d)\Bigg(p+\frac{\mu}{\mu+(i+1)\gamma}(s-p+(1-q)(r-s-lL(i,d)) \\ &+ h(i+1)\Bigg)+(1-f(d))h(i)\Bigg\}. &(100) \\ T_{DEP}h(i) &= h((i-1)^+). &(101) \\ T_{CAN}h(i) &= h((i-1)^+). &(102) \\ T_{FIC}h(i) &= h(i). &(103) \end{aligned}$$

We prove that the operators, T_{ARR}, T_{DEP}, T_{CAN}, T_{FIC} preserve the monotonicity properties of the function on which they are applied, i.e., $h(i)$. Actually, it is enough to show that T_{ARR} preserves the monotonicity in i, i.e., $T_{ARR}h(i) \geq T_{ARR}h(i+1)$ since the remaining operators are straightforward to show. Let d^0 and d^1 be the optimal lead-time quotes at states i and $i+1$, respectively. Then, since d^0 is optimal at state

i, we can write the following inequality comparing the values at d^0 and d^1.

$$\begin{aligned}
& f(d^0)\left(p+\frac{\mu}{\mu+(i+1)\gamma}(s-p+(1-q)(r-s-lL(i,d^0))+h^*(i+1)\right) \\
& +(1-f(d^0))h(i) \\
\geq\ & f(d^1)\left(p+\frac{\mu}{\mu+(i+1)\gamma}(s-p+(1-q)(r-s-lL(i,d^1))+h(i+1)\right) \\
& +(1-f(d^1))h(i). \qquad (104)
\end{aligned}$$

In order to obtain an inequality comparing $T_{ARR}h(i)$ and $T_{ARR}h(i+1)$, let us compare the right-hand side of Equation (104) with $T_{ARR}h(i+1)$.

$$\begin{aligned}
& f(d^1)\left(p+\frac{\mu}{\mu+(i+1)\gamma}(s-p+(1-q)(r-s-lL(i,d^1))+h(i+1)\right) \\
& +(1-f(d^1))h(i) \\
\geq\ & f(d^1)\left(p+\frac{\mu}{\mu+(i+2)\gamma}(s-p+(1-q)(r-s-lL(i+1,d^1))+h(i+2)\right) \\
& +(1-f(d^1))h(i+1), \qquad (105)
\end{aligned}$$

holds since $L(i,d)$ is strictly increasing in i by Lemma 6, and $\frac{\mu}{\mu+(i+1)\gamma} > \frac{\mu}{\mu+(i+2)\gamma}$.

Combining Equations (104) and (105), we prove that Equation (106) holds.

$$\begin{aligned}
& f(d^0)\left(p+\frac{\mu}{\mu+(i+1)\gamma}(s-p+(1-q)(r-s-lL(i,d^0))+h(i+1)\right) \\
& +(1-f(d^0))h(i) \\
\geq\ & f(d^1)\left(p+\frac{\mu}{\mu+(i+2)\gamma}(s-p+(1-q)(r-s-lL(i+1,d^1))+h(i+2)\right) \\
& +(1-f(d^1))h(i+1). \qquad (106)
\end{aligned}$$

Hence, $h_{n+1}(i)$ is monotone in i given that h_n is monotone in i. By taking the limit as $n \to \infty$, we obtain $\lim_{n\to\infty} h_n(i) = h^*(i)$ where $h^*(i)$ is the total profit of the system starting in state i in the infinite horizon version of the problem. Therefore structural results obtained by $h_n(i)$ hold for $h^*(i)$. Note that, Equation (99) can be viewed as the DP algorithm for a finite horizon problem with terminal profit function equal to $h_0(i)$ by reversing the time index; and $h_n(i)$ is the optimal profit starting from state i of a n-stage problem with $h_0(i)$ as terminal profit at the end of n-stages as in Bertsekas [4]. Moreover, these structural results are also valid for the average reward criteria version of the problem with relative value functions $h^*(i)$. $\square$

Proof of Proposition 3. Let us first define $\Pi_i = \max_d \left\{ \frac{\lambda}{\nu} f(d) \left(p + \frac{\mu}{\mu+(i+1)\gamma} (s - p + (1-q)(r - s - lL(i,d)) + h^*(i+1) - h^*(i) \right) \right\}$. Assume $d^*(i) \notin \Psi_i$ and $d^*(i) \neq d^{\max}$ by contradiction. Then, $\Pi_i^*(d(i)) < 0$ since by Lemma 7 $h^*(i+1) - h^*(i) \leq 0$. However, a natural lower bound on $\Pi_i^*(d(i))$ is zero by setting $d^* = d^{\max}$. $\square$

Proof of Theorem 2. We show that as p increases, g^* increases using the finite horizon version of the problem and the operators given in the proof of Lemma 7. The proof for r and s is similar, and hence omitted. Assume $p' > p$, and the relative value function for the case with p' is denoted by $v_n(i)$. We compare the relative value functions starting from $n = 1$ assuming $h_n(i) = 0$ for all i.

$$h_1^*(i) = \max_d \left\{ \frac{\lambda}{\nu} f(d) \left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma} p + \frac{\mu}{\mu+(i+1)\gamma} (qs + (1-q)(r - lL(i,d))) \right) \right\}. \tag{107}$$

$$v_1^*(i) = \max_d \left\{ \frac{\lambda}{\nu} f(d) \left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma} p' + \frac{\mu}{\mu+(i+1)\gamma} (qs + (1-q)(r - lL(i,d))) \right) \right\}. \tag{108}$$

Let d^0 be the maximizer of $h_1(i)$ and d^1 be the maximizer of $v_1(i)$. Then Equations (107)-(108) reduce to the following:

$$h_1^*(i) = \frac{\lambda}{\nu} f(d^0) \left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma} p + \frac{\mu}{\mu+(i+1)\gamma} (qs + (1-q)(r - lL(i,d^0))) \right). \tag{109}$$

$$v_1^*(i) = \frac{\lambda}{\nu} f(d^1) \left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma} p' + \frac{\mu}{\mu+(i+1)\gamma} (qs + (1-q)(r - lL(i,d^1))) \right). \tag{110}$$

Replacing d^1 in Equation (110) by d^0, we write the following inequality since d^1 is the maximizer.

$$\begin{aligned} & \frac{\lambda}{\nu} f(d^1) \left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma} p' + \frac{\mu}{\mu+(i+1)\gamma} (qs + (1-q)(r - lL(i,d^1))) \right) \\ \geq \; & \frac{\lambda}{\nu} f(d^0) \left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma} p' + \frac{\mu}{\mu+(i+1)\gamma} (qs + (1-q)(r - lL(i,d^0))) \right). \end{aligned} \tag{111}$$

We can also compare the right-hand side of Equation (111) and Equation (109) since $p' > p$:

$$\begin{aligned}
&\frac{\lambda}{\nu}f(d^0)\left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma}p' + \frac{\mu}{\mu+(i+1)\gamma}(qs+(1-q)(r-lL(i,d^0)))\right) \\
\geq\ &\frac{\lambda}{\nu}f(d^0)\left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma}p + \frac{\mu}{\mu+(i+1)\gamma}(qs+(1-q)(r-lL(i,d^0)))\right). \quad (112)
\end{aligned}$$

Now by combining Equations (111) and (112), we obtain the following equation and conclude that $v_1^*(i) \geq h_1^*(i)$.

$$\begin{aligned}
&\frac{\lambda}{\nu}f(d^1)\left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma}p' + \frac{\mu}{\mu+(i+1)\gamma}(qs+(1-q)(r-lL(i,d^1)))\right) \\
\geq\ &\frac{\lambda}{\nu}f(d^0)\left(\frac{(i+1)\gamma}{\mu+(i+1)\gamma}p + \frac{\mu}{\mu+(i+1)\gamma}(qs+(1-q)(r-lL(i,d^0)))\right). \quad (113)
\end{aligned}$$

Then by induction and same reasoning as in the proof of Lemma 7; we can conclude that as p increases, g^* increases. □

Proof of Theorem 3. The proof for this theorem is similar to the proof of Theorem 2, and hence omitted. □

APPENDIX C

SEQUENCE-BASED MIP FORMULATION FOR CHAPTER 4

This formulation is based on the fact that in the resulting solution of the problem all the tasks will be sequenced according to their start times. If we take any feasible solution to this problem and sort the tasks according to their start times, we will have a sequence of tasks in which tasks are visited by their required subset of workers. Since X is the set of tasks, tasks can be visited in positions $1, 2, ...|X|$. This visitation sequence may not give a visitation sequence for a particular worker since all tasks are not visited by every worker. Task visitation sequence for a given worker will not include the tasks that it is not required to visit. Excluding these tasks we can obtain task visitation sequence for any worker.

The decision variables used in the formulation are as follows: (i) the binary decision variables are x_{ij} which takes the value 1 if task i is visited in position j and 0 otherwise, (ii) D_{jk} are distance variables which reflect the distance between tasks in positions j and k, and (iii) T_j denotes the start time of task in position j. Our objective function in this problem is minimizing the total completion time, T which corresponds to start time of task in last position, $T_{|X|}$ in this formulation.

$$Minimize\ T \tag{114}$$

$$subject\ to$$

$$\sum_{j=1}^{|X|} x_{ij} = 1 \quad i = 1, ...|X| \tag{115}$$

$$\sum_{i=1}^{|X|} x_{ij} = 1 \quad j = 1, ...|X| \tag{116}$$

$$D_{jk} \geq (x_{ij} + x_{lk} - 1)d_{il}$$

$$j = 1, 2, ..., |X| - 1, \quad k = j + 1, ...|X| \quad R^i \cap R^l \neq \emptyset \quad i, l \in R \tag{117}$$

$$D_{0k} \geq x_{lk}d_{rl} \quad k = 1, 2, ...|X| \quad l \in X^r \tag{118}$$

$$T_k \geq T_j + D_{jk} \quad k = 1, 2, ..|X| \quad j = 0, 1, ...(k-1) \tag{119}$$

$$T_0 = 0 \tag{120}$$

$$T \geq T_k \quad k = 0, 1, 2, ...|X| \tag{121}$$

$$x_{ij} \in \{0, 1\} \quad i, j = 1, ...|X| \tag{122}$$

$$D_{jk} \geq 0 \quad j, k = 0, 1, ...|X| \tag{123}$$

$$T_j \geq 0 \quad j = 0, 1, ...|X| \tag{124}$$

In this formulation, (114) is the objective function which is the minimization of total completion time of all tasks. Constraints (115) ensure that each task is assigned to one position, and constraints (116) guarantee that each position is assigned to one of tasks. Constraints (117) calculate the distance between positions, i.e., the distance between tasks assigned to these positions. These constraints are active when the tasks in corresponding positions share at least one worker, otherwise it means that these positions are actually independent in sequence, and distance between tasks in these positions does not affect the distance between these positions since there are no workers visiting both of these tasks. Constraints (118) determine the distance to any position from worker's initial place. The start time of a task in position k must be larger than the sum of start time of a task in position j (for $j \leq k$) and distance between these positions as calculated. This is satisfied by constraints (119). (120) initializes the start time of workers in position 0 to 0. All workers are assigned to position 0 since all workers start at their initial place. Constraints (121) ensure that the total completion time of all tasks is greater than start time of tasks in all positions. Constraints (122), (123) and (124) are integrality and non-negativity constraints.

REFERENCES

[1] Ansberry, C. and Aeppel, T., "Battling imports: Surviving the onslaught; u.s. companies customize, rethink strategies to compete with products from abroad," *The Wall Street Journal*, October 6 2003.

[2] Ata, B., "Dynamic control of a multiclass queue with thin arrival streams," *Operations Research*, vol. 54, no. 5, pp. 876–892, 2006.

[3] Ata, B. and Olsen, T. L., "Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs," *Operations Research*, vol. 57, no. 3, pp. 753–768, 2009.

[4] Bertsekas, D. P., *Dynamic Programming and Optimal Control*, vol. 1. Belmont, Mass.: Athena Scientific, 3rd, ed., 2005.

[5] Bitran, G., Caldentey, R., and Vial, R., "Pricing policies for perishable products with demand substitution," *Tech. rep., Stern School of Business*, 2006.

[6] Bredstrom, D. and Ronnqvist, M., "Combined vehicle routing and scheduling with temporal precedence and synchronization constraints," *Department of Finance and Management Science, Norwegian School of Economocs and Business Administration (NHH) Norway*, 2006. Discussion Paper 2006/18.

[7] Ceryan, O., Sahin, O., and Duenyas, I., "Managing demand and supply for multiple products through dynamic pricing and flexibility," *Working paper*, 2011.

[8] Charnsirisakskul, K., Griffin, P. M., and Keskinocak, P., "Order selection and scheduling with leadtime flexibility," *IIE Transactions*, vol. 36, no. 7, pp. 697–707, 2004.

[9] Charnsirisakskul, K., Griffin, P. M., and Keskinocak, P., "Pricing and scheduling decisions with leadtime flexibility," *European Journal of Operational Research*, vol. 171, no. 1, pp. 153–169, 2006.

[10] Cheung, K. L. and Zhang, A. X., "The impact of inventory information distortion due to customer order cancellations," *Naval Research Logistics*, vol. 46, pp. 213–231, MAR 1999.

[11] Cil, E. B., "General framework to perform sensitivity analysis on markovian queueing and inventory systems.," Master's thesis, Koc University, 2006.

[12] Cohen, M. A., Teck, H. H., Ren, Z. J., and Terwiesch, C., "Measuring imputed cost in the semiconductor equipment supply chain," *Management Science*, vol. 49, pp. 1653–1670, December 2003.

[13] Dohn, A., Kolind, E., and Clausen, J., "The manpower allocation problem with time windows and job-teaming constraints: A branch and price approach," *Computers & Operations Research*, vol. 36, pp. 1145–1157, 2009.

[14] Dong, L., Kouvelis, P., and Tian, Z., "Dynamic pricing and inventory control of substitute products," *Manufacturing & Service Operations Management*, vol. 11, no. 2, pp. 317–339, 2009.

[15] Duenyas, I., "Single facility due date setting with multiple customer classes," *Management Science*, vol. 41, no. 4, pp. 608–619, 1995.

[16] Duenyas, I. and Hopp, W. J., "Quoting customer lead times," *Management Science*, vol. 41, no. 1, pp. 43–57, 1995.

[17] Economist, "A long march. mass customization.," *Economist*, vol. 360, p. 8230, 2001.

[18] Ekici, A., Keskinocak, P., and Koenig, S., "Multi-robot routing with linear decreasing rewards over time," *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, pp. 958–963, 2009.

[19] Eskew, M., "The art of delivery: Why more ceos are becoming supply chain experts," *The Chief Executive*, vol. August-Sept, 2004.

[20] Frame, P., "Saturn to fine suppliers $500/minute for delays," *Automotive News*, vol. 21, December 1992.

[21] Gaur, V. and Honhon, D., "Assorment planning and inventory decisions under a locational choice model," *Management Science*, vol. 52, no. 10, pp. 1528–1543, 2006.

[22] Gayon, J. P., Bensaafa, r. S., and Vericourt, F., "Using imperfect advance demand information in production-inventory systems with multiple customer classes," *Manufacturing and Service Operations Management*, vol. 11, no. 1, pp. 128–143, 2009.

[23] George, F., "The new aircraft sales contract," *Business and Commercial Aviation*, vol. 86, no. 3, pp. 90–93, 2000.

[24] Gerkey, B. P. and Mataric, M. J., "A formal analysis and taxonomy of task allocation in multi-robot systems," *International Journal of Robotics Research*, vol. 23, no. 9, pp. 939–954, 2004.

[25] Gunasekaran, A., Patel, C., and McGaughey, R. E., "A framework for supply chain performance measurement," *Int. J. Production Economics*, vol. 87, no. 10, pp. 333–347, 2004.

[26] Hafizoglu, A. B., Gel, E. S., and Keskinocak, P., "Expected tardiness computations in multi class priority m/m/c queues," *Informs Journal on Computing*, 2012.

[27] Heyman, D. P. and Sobel, M. J., *Stochastic Models in Operations Research*, vol. 1. Dover, 2004.

[28] Jonhson, D. S. and Papadimitriou, C. H., *Performance Guarantees for Heuristics. In The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Lawler, E. L., J. K. Lenstra, A. H. G. Rinnooy Kan and D. B. Shmoys (eds.), John Wiley & Sons Ltd., New York, 1985.

[29] Kaminsky, P. and Hochbaum, D., *Due date quotation models and algorithms*. Boca Raton, FL.: CRC Press, j. y.-t. leung. ed. handbook of scheduling: algoritgms, models, and performance analysis ed., 2004.

[30] Kapuscinski, R. and Tayur, S., "Reliable due-date setting in a capacitated mto system with two customer classes," *Operations Research*, vol. 55, no. 1, pp. 56–74, 2007.

[31] Keskinocak, P., Ravi, R., and Tayur, S., "Scheduling and reliable lead-time quotation of orders with availability intervals and lead-time sensitive revenues," *Management Science*, vol. 47, no. 2, pp. 264–279, 2001.

[32] Keskinocak, P. and Tayur, S., *Due-date Management Policies*. Kluwer Academic Publisher, 2004.

[33] Li, L., "The role of inventory in delivery-time competition," *Management Science*, vol. 38, no. 2, pp. 182–197, 1992.

[34] Li, Y., Lim, A., and Rodrigues, B., "Manpower allocation with time windows and job-teaming constraints," *Naval Research Logistics*, vol. 52, pp. 302–311, 2005.

[35] Lim, A., Rodrigues, B., and Song, L., "Manpower allocation with time windows," *Journal of the Operational Research Society*, vol. 55, pp. 1178–1186, 2004.

[36] Lippman, S. A., "Applying a new device in the optimization of exponential queuing systems," *Operations Research*, vol. 23, no. 4, pp. 687–710, 1965.

[37] Lundh, R., Karlsson, L., and Saffiotti, A., "Autonomous functional configuration of a network robot system," *Robotics and Autonomous Systems*, vol. 56, no. 10, pp. 819–830, 2008.

[38] Maddah, B. and Bish, E. K., "Joint pricing, assortment and inventory decisions for a retailer's product line," *Operations Research*, vol. 54, pp. 315–330, 2007.

[39] Mahajan, R. and Ryzin, G. V., "Stocking retail assortments under dynamic consumer substitution," *Operations Research*, vol. 49, no. 3, pp. 334–351, 2001.

[40] Monnot, J., "Approximation algorithms for the maximum hamiltonian path problem with specified endpoint(s)," *European Journal of Operational Research*, vol. 161, pp. 721–735, MAR 16 2005.

[41] Ostrower, J., "Boeing 787 profit plan strained by order cut," *The Wall Street Journal*, August 23, 2012.

[42] Parker, L. E. and Tang, F., "Building multirobot coalitions through automated task solution synthesis," *Proceedings of the IEEE (Special Issue on Multi-Robot Systems)*, vol. 94, no. 7, pp. 1289–1305, 2006.

[43] Pekgun, P., Griffin, P. M., and Keskinocak, P., "Centralized vs. decentralized competition for price and lead-time sensitive demand," *submitted to Manufacturing and Service Operations Management*, 2009.

[44] Plambeck, E. L., "Optimal leadtime differentiation via diffusion approximations," *Operations Research*, vol. 52, no. 2, pp. 213–228, 2004.

[45] Ross, S. M., *Introduction to Probability Models*. Academic Press, 2007.

[46] Rubino, M. and Ata, B., "Dynamic control of a make-to-order, parallel-server system with cancellations," *Operations Research*, vol. 57, pp. 94–108, JAN-FEB 2009.

[47] Russell, R. and Taylor, B., *Operations Management: Focusing on Qulality and Competitiveness*. Prentice-Hall, 1998.

[48] Savasaneril, S., Griffin, P. M., and Keskinocak, P., "Dynamic lead-time quotation for an m/m/1 base-stock inventory queue," *Operations Research*, vol. 58, no. 2, pp. 383–395, 2010.

[49] Slotnick, S. A. and Sobel, M. J., "Manufacturing lead-time rules: Customer retention versus tardiness costs," *Eur. J. Oper. Res.*, vol. 163, no. 3, pp. 825–826, 2005.

[50] Smith, S. A. and Agrawal, N., "Management of multi-item retail inventory systems with demand substitution," *Operations Research*, vol. 48, no. 1, pp. 50–64, 2000.

[51] So, K. C. and Zheng, X., "Impact of supplier's lead time and forecast demand updating on retailer's order quantity variability in a two-level supply chain," *Int. J. Production Economics*, vol. 86, pp. 169–179, 2003.

[52] Suh, M. and Aydin, G., "Dynamic pricing of substitutable products with limited inventories under logit demand," *IIE Transactions*, vol. 43, no. 5, pp. 323–331, 2011.

[53] van Ryzin, G. and Mahajan, S., "On the relationship between inventory costs and variety benefits in retail assortments," *Management Science*, vol. 45, no. 11, pp. 1496–1509, 1999.

[54] Walpole, R. E., Myers, R., Myers, S., and Ye, K., *Probability and Statistics For Engineers and Scientists*. Upper Saddle River, New Jersey: Prentice-Hall, 2002.

[55] Yeo, W. M. and Yuan, X. M., "Optimal inventory policy with supply uncertainty and demand cancellation," *European Journal of Operational Research*, vol. 211, pp. 26–34, MAY 16 2011.

[56] You, P.-S., "Dynamic pricing of inventory with cancellation demand," *The Journal of the Operational Research Society*, vol. 54, pp. 1093–1101, Oct., 2003 2003.

[57] You, P. S. and Hsieh, Y. C., "A lot size model for deteriorating inventory with back-order cancellation," *Springer Lecture Notes in Computer Science*, vol. 4570, pp. 1002–1011, 2007.

[58] You, P. S. and Wu, M. T., "Optimal ordering and pricing policy for an inventory system with order cancellations," *OR Spectrum*, vol. 29, pp. 661–679, 2007.

[59] Yuan, X. M. and Cheung, K. L., "Optimal inventory policy for systems with demand cancellations," *IIE Transactions*, vol. 35, pp. 739–747, AUG 2003.

[60] Zheng, X. and Koenig, S., "Greedy approaches for solving task-allocation problems with coalitions," *AAMAS 2008 Workshop on Formal Models and Methods for Multi-Robot Systems*, pp. 35–40, 2008.

[61] Zheng, X. and Koenig, S., "Reaction functions for task allocation to cooperative agents," *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 559–566, 2008.